

# Nonparametric Bayesian Negative Binomial Factor Analysis

Mingyuan Zhou

The University of Texas at Austin

January 25, 2016

## Abstract

A common approach to analyze an attribute-instance count matrix, an element of which represents how many times an attribute appears in an instance, is to factorize it under the Poisson likelihood. We show its limitation in capturing the tendency for an attribute present in an instance to both repeat itself and excite related ones. To address this limitation, we construct negative binomial factor analysis (NBFA) to factorize the matrix under the negative binomial likelihood, and relate it to a Dirichlet-multinomial distribution based mixed-membership model. To support countably infinite factors, we propose the hierarchical gamma-negative binomial process. By exploiting newly proved connections between discrete distributions, we construct two blocked and a collapsed Gibbs sampler that all adaptively truncate their number of factors, and demonstrate that the blocked Gibbs sampler developed under a compound Poisson representation converges fast and has low computational complexity. Example results show that NBFA has a distinct mechanism in adjusting its number of inferred factors according to the instance lengths, and provides clear advantages in parsimonious representation, predictive power, and computational complexity over previously proposed discrete latent variable models, which either completely ignore burstiness, or model only the burstiness of the attributes but not that of the factors.

*Keywords:* burstiness, count matrix factorization, hierarchical gamma-negative binomial process, parsimonious representation, self- and cross-excitation.

---

M. Zhou is an Assistant Professor of Statistics in the Department of Information, Risk, & Operations Management and the Department of Statistics & Data Sciences at The University of Texas at Austin, Austin, TX 78712, USA. *Email:* mingyuan.zhou@mcombs.utexas.edu

# 1 Introduction

The need to analyze an attribute-instance count matrix, each of whose elements counts the number of time that an attribute appears in an instance, arises in many different settings, such as text analysis, next-generation sequencing, medical records mining, and consumer behavior studies. The mixed-membership model, independently developed for text analysis (Blei et al., 2003) and population genetics (Pritchard et al., 2000), treats each instance as a bag of tokens, and associates each token with both the index of an attribute that is observed and the index of a subpopulation that is latent. It makes the assumption that there are  $K$  latent subpopulations, each of which is characterized by how frequent the attributes are relative to each other in that subpopulation. Given the total number of tokens for an instance, it assigns each token independently to one of the  $K$  subpopulations, with a probability proportional to the product of the corresponding attribute’s relative frequency in that subpopulation and that subpopulation’s relative frequency in the instance. A mixed-membership model constructed in this manner, as shown in Zhou et al. (2012) and Zhou and Carin (2015), can also be connected to Poisson factor analysis (PFA) that factorizes the attribute-instance count matrix, under the Poisson likelihood, into the product of a nonnegative attribute-subpopulation factor loading matrix and a nonnegative subpopulation-instance factor score matrix. Each column of the factor loading matrix encodes the relative frequencies of the attributes in a subpopulation, while each column of the factor score matrix encodes the weights of the subpopulations in an instance.

Despite the popularity of both approaches in analyzing the attribute-instance count matrix, they both make restrictive assumptions. Given the relative frequencies of attributes in subpopulations and the relative frequencies of subpopulations in an instance, the mixed-membership model generates both the attribute and subpopulation indices of a token independently from these of the other tokens, and hence may not sufficiently capture the tendency for a token to excite the other ones in the same instance to take the same or related attributes. Whereas for PFA, given the factor loading and score matrices, it assumes

that the variance and mean are the same for each attribute-instance count, and hence is likely to underestimate the variability of overdispersed counts.

In practice, however, highly overdispersed attribute-instance counts are frequently observed due to self- and cross-excitation of attribute frequencies. For example, the tendency for an attribute present in a document to appear repeatedly is a fundamental phenomenon in natural language that is commonly referred to as word burstiness (Church and Gale, 1995; Madsen et al., 2005; Doyle and Elkan, 2009). If a word is bursty in a document, it is also common for it to excite related words to exhibit burstiness. Without capturing the self- and cross-excitation of attribute frequencies or better modeling the overdispersed attribute-instance counts, the ultimate potential of the mixed-membership model and PFA will be limited no matter how the priors on latent parameters are adjusted. In addition, it could be a waste of computation if the model tries to increase the model capacity to better capture overdispersions that could be simply explained with self- and cross-excitations.

To remove these restrictions, we introduce negative binomial factor analysis (NBFA) to factorize the attribute-instance count matrix, in which we replace the Poisson distributions on which PFA is built, with the negative binomial (NB) distributions. As PFA is closely related to the canonical mixed-membership model built on the multinomial distribution, we show that NBFA is closely related to a Dirichlet-multinomial mixed-membership (DMMM) model that uses the Dirichlet-categorical (Dirichlet-multinomial) rather than categorical (multinomial) distributions to generate both the attribute and factor (subpopulation) indices of the tokens. From the viewpoint of count modeling, NBFA improves PFA by better modeling overdispersed counts, while from that of mixed-membership modeling, the DMMM model improves the canonical mixed-membership model by capturing the burstiness of both the attribute and factor indices. In addition, we will show NBFA could significantly reduce the computation spent on large attribute-instance counts.

## 1.1 Related algorithms

To make connections to Dirichlet compound multinomial latent Dirichlet allocation (DCMLDA) of Doyle and Elkan (2009), which improves the canonical mixed-membership model by modeling the burstiness of the attributes, we show DCMLDA can be considered as a simplified DMMM model that models attribute burstiness but not factor burstiness. We further relate DCMLDA to NBFA by restricting the factor scores of NBFA to be the same for all instances, under the same set of factors shared by all instances.

Note that with a different likelihood for counts and a different mechanism to generate both the attribute and factor indices, NBFA and the DMMM model proposed in the paper complement, rather than compete with, PFA (Zhou et al., 2012; Zhou and Carin, 2015). First, NBFA provides more significant advantages in modeling longer instances, where there is more need to capture both self- and cross-excitation of attribute frequencies. Second, various extensions built on PFA or the multinomial mixed-membership model, such as imposing different priors on the subpopulation proportions (Wallach et al., 2009) or unnormalized factor scores (Zhou and Carin, 2015), capturing the correlations between factors (Blei and Lafferty, 2006a; Paisley et al., 2012; Ranganath and Blei, 2015), modeling the temporal evolutions of subpopulation proportions (Blei and Lafferty, 2006b) or unnormalized factor scores (Acharya et al., 2015), arranging factors under a tree structure (Blei et al., 2010; Adams et al., 2010; Paisley et al., 2015), and learning multilayer deep representations (Ranganath et al., 2015; Gan et al., 2015; Zhou et al., 2015), could also be applied to extend NBFA. In this paper, we will focus on constructing a nonparametric Bayesian NBFA with a potentially infinite number of factors, and leave a wide variety of potential extensions under this new framework to future research.

## 1.2 Nonparametric Bayesian modeling

To avoid the need of selecting the number of subpopulations  $K$ , for PFA and the closely related multinomial mixed-membership model, a number of nonparametric Bayesian priors can be employed to support a potentially infinite number of latent factors, including the gamma-

Poisson process (Lo, 1982; Titsias, 2008; Zhou and Carin, 2015), hierarchical Dirichlet process (Teh et al., 2006), Indian buffet process (Griffiths and Ghahramani, 2005; Williamson et al., 2010), beta-negative binomial process (Zhou et al., 2012; Broderick et al., 2015; Zhou and Carin, 2015; Heaukulani and Roy, 2015; Zhou et al., 2016), and gamma-negative binomial process (Zhou and Carin, 2015; Zhou et al., 2016).

To support countably infinite factors for NBFA, we introduce a new nonparametric Bayesian prior: the hierarchical gamma-negative binomial process (hGNBP), where each of the  $J$  instances is assigned with an instance-specific gamma-negative binomial process (GNBP) and a globally shared gamma process is mixed with all the  $J$  GNBP. We derive both blocked and collapsed Gibbs sampling for the hGNBP-NBFA, with the number of factors automatically inferred.

To make connections to DCMLDA, we also consider NBFA based on the GNBP, in which each of the  $J$  instances is assigned with an instance-specific NB process and a globally shared gamma process is mixed with all the  $J$  NB processes. We call this nonparametric Bayesian model as the GNBP-DCMLDA, which is shown to be restrictive in that although each instance has its own factor scores under the corresponding instance-specific factors, all the instances are enforced to have the same factor scores under the same set of globally shared factors. By contrast, by modeling not only the burstiness of the attributes, but also that of the factors, the hGNBP-NBFA provides instance-specific factor scores under the same set of shared factors, making it suitable for extracting low-dimensional latent representations for high-dimensional attribute count vectors.

The remainder of the paper is organized as follows. In Section 2, we review some useful discrete distributions and PFA. In Section 3, we introduce NBFA and its representation as a DMMM model, and compare them with related models. In Section 4, we propose nonparametric-Bayesian NBFA using the hGNBP, and derive both blocked and collapsed Gibbs sampling algorithms. In Section 5, we first make comparisons between different sampling strategies and then compare the performance of various algorithms on real data. We

conclude the paper in Section 6. The proofs and Gibbs sampling update equations are provided in the Appendix.

## 2 Preliminaries

### 2.1 Review on some useful discrete distributions

Denote  $\mathbf{b} \sim \text{CRP}(n, r)$  as an exchangeable random partition of the set  $\{1, \dots, n\}$ , generated by assigning  $n$  customers (instances) to  $\ell$  random number of tables (partitions) using a Chinese restaurant process (CRP) with concentration parameter  $r$ . The exchangeable partition probability function (EPPF) of  $\mathbf{b}$  under the CRP, also known as the Ewens sampling formula (Ewens, 1972; Antoniak, 1974), can be expressed as  $P(\mathbf{b} | n, r) = \frac{\Gamma(r)r^\ell}{\Gamma(n+r)} \prod_{t=1}^{\ell} \Gamma(n_t)$ , where  $n_t = \sum_{i=1}^n \delta(b_i = t)$  is the size of partition  $t$  and  $\ell$  is the number of partitions (Pitman, 2006). Denote  $\ell \sim \text{CRT}(n, r)$  as the Chinese restaurant table (CRT) random variable generated as the summation of  $n$  independent Bernoulli random variables as  $\ell = \sum_{i=1}^n b_i$ ,  $b_i \sim \text{Bernoulli}(\frac{r}{r+i-1})$ . The probability mass function (PMF) of  $\ell \sim \text{CRT}(n, r)$  can be expressed as  $f_L(\ell | n, r) = \frac{\Gamma(r)r^\ell}{\Gamma(n+r)} |s(n, \ell)|$ , where  $\ell \in \{0, 1, \dots, n\}$  and  $|s(n, \ell)| = \frac{n!}{\ell!} \sum_{(n_1, \dots, n_\ell) \in \mathcal{D}_{n, \ell}} \prod_{t=1}^{\ell} \frac{1}{n_t!}$  are unsigned Stirling numbers of the first kind (Johnson et al., 1997) that can be obtained by summing over the elements of the set  $\mathcal{D}_{n, \ell} = \{(n_1, \dots, n_\ell) : n_t \geq 1 \text{ and } \sum_{t=1}^{\ell} n_t = n\}$ .

Let  $n \sim \text{NB}(r, p)$  denote the NB distribution with PMF  $f_N(n | r, p) = \frac{\Gamma(n+r)}{n! \Gamma(r)} p^n (1-p)^r$ , where  $n \in \{0, 1, \dots\}$ , and let  $u \sim \text{Log}(p)$  denote the logarithmic distribution (Fisher et al., 1943) with PMF  $f_U(u | p) = \frac{1}{-\ln(1-p)} \frac{p^u}{u}$ , where  $u \in \{1, 2, \dots\}$ . Denote  $x \sim \text{SumLog}(\ell, p)$  as the sum-logarithmic random variable (Zhou et al., 2016) generated as  $x = \sum_{t=1}^{\ell} u_t$ ,  $u_t \sim \text{Log}(p)$ , with PMF  $f_N(n | \ell, p) = \frac{p^n \ell! |s(n, \ell)|}{n! [-\ln(1-p)]^\ell}$ . As revealed in Zhou and Carin (2015), the joint distribution of  $n$  and  $\ell$  given  $r$  and  $p$  in  $\ell | n \sim \text{CRT}(n, r)$ ,  $n \sim \text{NB}(r, p)$  is the same as that in  $n | \ell \sim \text{SumLog}(\ell, p)$ ,  $\ell \sim \text{Pois}[-r \ln(1-p)]$ , which is called as the Poisson-logarithmic bivariate distribution, with PMF

$$f_{N,L}(n, \ell | r, p) = \frac{|s(n, \ell)| r^\ell}{n!} p^n (1-p)^r. \quad (1)$$

As in Mosimann (1962) and Johnson et al. (1997), marginalizing out  $\boldsymbol{\theta}$  from  $\mathbf{z} \sim \prod_{i=1}^n \text{Cat}(z_i; \boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \sim \text{Dir}(r_1, \dots, r_K)$  leads to a Dirichlet-categorical (DirCat) distribution with PMF

$$P(\mathbf{z} | r_1, \dots, r_K) = \frac{\Gamma(r_{\cdot})}{\Gamma(n + r_{\cdot})} \prod_{k=1}^K \frac{\Gamma(n_k + r_k)}{\Gamma(r_k)}, \quad (2)$$

where  $n_k = \sum_{i=1}^n \delta(z_i = k)$ , and a Dirichlet-multinomial (DirMult) distribution with PMF  $P[(n_1, \dots, n_K) | r_1, \dots, r_K] = \frac{n!}{\prod_{k=1}^K n_k!} P(\mathbf{z} | r_1, \dots, r_K)$ .

## 2.2 Poisson factor analysis and mixed-membership model

PFA factorizes the attribute-instance count matrix under the Poisson likelihood as

$$\mathbf{N} \sim \text{Pois}(\boldsymbol{\Phi} \boldsymbol{\Theta}), \quad (3)$$

where  $\boldsymbol{\Phi} = (\phi_1, \dots, \phi_K) \in \mathbb{R}_+^{V \times K}$  represents the factor loading matrix,  $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J) \in \mathbb{R}_+^{K \times J}$  represents the factor score matrix, and  $\mathbb{R}_+ = \{x : x \geq 0\}$ , with  $\phi_k = (\phi_{1k}, \dots, \phi_{V_k})^T$  encoding the weights of the  $V$  attributes in factor  $k$  and  $\boldsymbol{\theta}_j = (\theta_{1j}, \dots, \theta_{Kj})^T$  encoding the popularity of the  $K$  factors in instance  $j$ . PFA in (3) has an augmented representation as

$$n_{vj} = \sum_{k=1}^K n_{vjk}, \quad n_{vjk} \sim \text{Pois}(\phi_{vk} \theta_{kj}). \quad (4)$$

As in Zhou et al. (2012), it can also be equivalently constructed by first generating  $n_{vj}$  and then assigning them to the latent factors using the multinomial distributions as

$$(n_{vj1}, \dots, n_{vjK}) | n_{vj} \sim \text{Mult}\left(n_{vj}, \frac{\phi_{v1} \theta_{1j}}{\sum_{k=1}^K \phi_{vk} \theta_{kj}}, \dots, \frac{\phi_{vK} \theta_{Kj}}{\sum_{k=1}^K \phi_{vk} \theta_{kj}}\right), \quad n_{vj} \sim \text{Pois}\left(\sum_{k=1}^K \phi_{vk} \theta_{kj}\right). \quad (5)$$

This alternative representation suggests a potential link of PFA to a standard mixed-membership model for text analysis such as probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) and latent Dirichlet allocation (LDA) (Blei et al., 2003). Given the factors

$\phi_k$  and factor proportions  $\theta_j/\theta_{\cdot j}$ , where  $\sum_{v=1}^V \phi_{vk} = 1$  and  $\theta_{\cdot j} := \sum_{k=1}^K \theta_{kj}$ , a standard procedure is to associate  $x_{ji} \in \{1, \dots, V\}$ , the  $i$ th attribute (word) token of the  $j$ th instance (document), with factor (topic)  $z_{ji} \in \{1, \dots, K\}$ , and generate a bag of attribute (word) tokens  $\{x_{j1}, \dots, x_{jn_j}\}$  as

$$x_{ji} \sim \text{Cat}(\phi_{z_{ji}}), \quad z_{ji} \sim \text{Cat}(\theta_j/\theta_{\cdot j}), \quad (6)$$

where  $n_j = \sum_{v=1}^V n_{vj}$  and  $n_{vj} = \sum_{i=1}^{n_j} \delta(x_{ji} = v)$ . We refer to (6) as the multinomial mixed-membership model. LDA completes the multinomial mixed-membership model by imposing the Dirichlet priors on both  $\{\phi_k\}_k$  and  $\{\theta_j/\theta_{\cdot j}\}_j$  (Blei et al., 2003).

If in addition to the multinomial mixed-membership model described in (6), one further generates the instance lengths as

$$n_j \sim \text{Pois}(\theta_{\cdot j}), \quad (7)$$

then the joint likelihood of  $\mathbf{x}_j := (x_{j1}, \dots, x_{jn_j})$ ,  $\mathbf{z}_j := (z_{j1}, \dots, z_{jn_j})$ , and  $n_j$  given  $\Phi$  and  $\theta_j$  can be expressed as  $P(\mathbf{x}_j, \mathbf{z}_j, n_j | \Phi, \theta_j) = \frac{\prod_{v=1}^V \prod_{k=1}^K n_{vjk}!}{n_j!} \prod_{v=1}^V \prod_{k=1}^K \text{Pois}(n_{vjk}; \phi_{vk} \theta_{kj})$ , whose product with the combinatorial coefficient  $n_j! / (\prod_{v=1}^V \prod_{k=1}^K n_{vjk}!)$  becomes the same as the likelihood  $P(\{n_{vj}, n_{vj1}, \dots, n_{vjK}\}_v | \Phi, \theta_j)$  of (4).

From the viewpoint of PFA, shown in (4), and its alternative representation constituted by (6) and (7), a wide variety of discrete latent variable models, such as nonnegative matrix factorization (NMF) (Lee and Seung, 2001), PLSA, LDA, the gamma-Poisson model of Canny (2004), the discrete component analysis of Buntine and Jakulin (2006), and the correlated topic model (Blei and Lafferty, 2006a), all have the same mechanism to model the attribute counts that they generate both the attribute and factor indices using the categorical distributions shown in (6); they mainly differ from each other on how the priors on  $\phi_k$  and  $\theta_j$  (or  $\theta_j/\theta_{\cdot j}$ ) are constructed (Zhou et al., 2012; Zhou and Carin, 2015).



### 3 Negative binomial factor analysis and the Dirichlet-multinomial mixed-membership model

#### 3.1 Negative binomial factor analysis

To better model overdispersed counts, instead of following PFA to factorize the attribute-instance count matrix under the Poisson likelihood, we propose negative binomial factor analysis (NBFA) to factorize it under the NB likelihood as

$$\mathbf{n}_j \sim \text{NB}(\Phi \boldsymbol{\theta}_j, p_j). \quad (8)$$

NBFA has an augmented representation as

$$n_{vj} = \sum_{k=1}^K n_{vjk}, \quad n_{vjk} \sim \text{NB}(\phi_{vk} \theta_{kj}, p_j). \quad (9)$$

Similar to how the Poisson is related to the multinomial distribution (*e.g.*, Dunson and Herring (2005) and Lemma 4.1 of Zhou et al. (2012)), we reveal how the NB distribution is related to the Dirichlet-multinomial distribution using the following theorem, whose proof is provided in Appendix A.

**Theorem 1** (The negative binomial and Dirichlet-multinomial distributions). *Let  $\mathbf{x} = (x, x_1, \dots, x_K)$  be random variables generated as*

$$x = \sum_{k=1}^K x_k, \quad x_k \sim \text{NB}(r_k, p).$$

*Set  $r. = \sum_{k=1}^K r_k$  and let  $\mathbf{y} = (y, y_1, \dots, y_K)$  be random variables generated as*

$$(y_1, \dots, y_K) \sim \text{DirMult}(y, r_1, \dots, r_K), \quad y \sim \text{NB}(r., p).$$

*Then the distribution of  $\mathbf{x}$  is the same as that of  $\mathbf{y}$ .*

Using Theorem 1,  $(n_{vj}, n_{vj1}, \dots, n_{vjK})$  in (9) can be equivalently generated as

$$(n_{vj1}, \dots, n_{vjK}) | n_{vj} \sim \text{DirMult}(n_{vj}, \phi_{v1}\theta_{1j}, \dots, \phi_{vK}\theta_{Kj}), \quad n_{vj} \sim \text{NB}\left(\sum_{k=1}^K \phi_{vk}\theta_{kj}, p_j\right). \quad (10)$$

Clearly, how the factorization under the NB likelihood is related to the Dirichlet-multinomial distribution, as in (9) and (10), mimics how the factorization under the Poisson likelihood is related to the multinomial distribution, as in (4) and (5).

### 3.2 The Dirichlet-multinomial mixed-membership model

Similar to how we relate PFA in (4) to the multinomial topic model in (6), as discussed in Section 1, we may relate NBFA in (9) to a mixed-membership model constructed by replacing the categorical distributions in (6) with the Dirichlet-categorical distributions as

$$\begin{aligned} x_{ji} &\sim \text{Cat}(\phi_{z_{ji}}^{[j]}), \quad z_{ji} \sim \text{Cat}(\theta^{[j]}), \\ \phi_k^{[j]} &\sim \text{Dir}(\phi_k \theta_{kj}), \quad \theta^{[j]} \sim \text{Dir}(\theta_j), \end{aligned} \quad (11)$$

where  $\{\phi_k^{[j]}\}_k$  and  $\theta^{[j]}$  represent the factors and factor scores specific for instance  $j$ , respectively. Introducing  $\phi_k^{[j]}$  into the hierarchical model allows the same set of factors  $\{\phi_k\}_k$  to be manifested differently in different instances, whereas introducing  $\theta^{[j]}$  allows each instance to have two different representations: the factor scores  $\theta^{[j]}$  under the instance-specific factors  $\{\phi_k^{[j]}\}_k$ , and the factor scores  $\theta_j$  under the shared factors  $\{\phi_k\}_k$ . In addition, under this construction, the variance-to-mean ratio of  $\phi_{vk}^{[j]}$  given  $\phi_k$  and  $\theta_{kj}$  becomes  $(1 - \phi_{vk})/(\theta_{kj} + 1)$ , which monotonically decreases as the corresponding factor score  $\theta_{kj}$  increases, allowing the variability of  $\phi_k^{[j]}$  in the prior to be controlled by the popularity of  $\phi_k$  in the corresponding instance. Moreover, this construction helps simplify the model likelihood and allows the model to be closely related to NBFA, as discussed below.

Explicitly instantiating  $\{\phi_k^{[j]}\}_k$  for all instances would be computationally prohibitive, especially if the number of instances is large. Fortunately, this operation is totally unnecessary. By marginalizing out  $\phi_k^{[j]}$  and  $\theta^{[j]}$  in (11), we have

$$\{x_{ji}\}_{i:z_{ji}=k} \sim \text{DirCat}(n_{\cdot jk}, \phi_{1k}\theta_{kj}, \dots, \phi_{V_k}\theta_{kj}), \quad \mathbf{z}_j \sim \text{DirCat}(n_j, \boldsymbol{\theta}_j), \quad (12)$$

where  $\mathbf{z}_j := (z_{j1}, \dots, z_{jn_j})$ ,  $n_{vjk} := \sum_{i=1}^{n_j} \delta(x_{ji} = v, z_{ji} = k)$ , and  $n_{\cdot jk} := \sum_{v=1}^V n_{vjk}$ . Thus the joint likelihood of  $\mathbf{x}_j := (x_{j1}, \dots, x_{jn_j})$  and  $\mathbf{z}_j$  given  $\boldsymbol{\Phi}$ ,  $\boldsymbol{\theta}_j$ , and  $n_j$  can be expressed as

$$P(\mathbf{x}_j, \mathbf{z}_j \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, n_j) = \frac{\Gamma(\theta_{\cdot j})}{\Gamma(n_j + \theta_{\cdot j})} \prod_{v=1}^V \prod_{k=1}^K \frac{\Gamma(n_{vjk} + \phi_{vk}\theta_{kj})}{\Gamma(\phi_{vk}\theta_{kj})}. \quad (13)$$

We call the model shown in (11) or (12) as the Dirichlet-multinomial mixed-membership (DMMM) model, whose likelihood given the factors and factor scores is shown in (13).

We introduce the following proposition, with proof provided in Appendix A, to show that one can recover NBFA from the DMMM model by randomizing its instance lengths with NB distributions, and can reduce NBFA to the DMMM model by conditioning on these lengths.

**Proposition 2** (Dirichlet-multinomial mixed-membership modeling and negative binomial factor analysis). *For the Dirichlet-multinomial mixed-membership (DMMM) model that generates the attribute and factor indices using (11) or (12), if the instance lengths are randomized as*

$$n_j \sim \text{NB}(\theta_{\cdot j}, p_j), \quad (14)$$

*then the joint likelihood of  $\mathbf{x}_j$ ,  $\mathbf{z}_j$ , and  $n_j$  given  $\boldsymbol{\Phi}$ ,  $\boldsymbol{\theta}_j$ , and  $p_j$ , multiplied by the combinatorial coefficient  $n_j! / (\prod_{v=1}^V \prod_{k=1}^K n_{vjk}!)$ , is equal to the likelihood of negative binomial factor analysis (NBFA) described in (9) or (10), expressed as*

$$P(\{n_{vj1}, n_{vj2}, \dots, n_{vjK}\}_{v=1, V} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j) = \prod_{v=1}^V \prod_{k=1}^K \text{NB}(n_{vjk}; \phi_{vk}\theta_{kj}, p_j). \quad (15)$$

The DMMM model could model not only the burstiness of the attributes, but also that of the factors via the Dirichlet-categorical distributions, as further explained below when discussing related models. As far as the conditional posteriors of  $\boldsymbol{\phi}_k$  and  $\boldsymbol{\theta}_j$  are concerned,

the DMMM model with the lengths of its instances randomized via the NB distributions, as shown in (12) and (14), is equivalent to NBFA, as shown in (9). The differences in these two representations, however, lead to different inference strategies, which will be discussed in detail along with their nonparametric Bayesian generalizations.

### 3.3 Comparisons with related models

Preceding the DMMM model proposed in this paper, to account for attribute burstiness, Doyle and Elkan (2009) proposed Dirichlet compound multinomial LDA (DCMLDA) that can be expressed as

$$x_{ji} \sim \text{Cat}(\phi_{z_{ji}}^{[j]}), \quad z_{ji} \sim \text{Cat}(\theta_j), \quad \phi_k^{[j]} \sim \text{Dir}(\phi_k r_k), \quad (16)$$

where the Dirichlet prior is further imposed on  $\theta_j$ . Note that the instance-specific factor scores  $\{\theta_j\}_j$  are represented under the instance-specific factors  $\{\phi_k^{[j]}\}_k$  in DCMLDA, as shown in (16), whereas they are represented under the same set of factors  $\{\phi_k\}_k$  in the DMMM model, as shown in (12).

Comparing (11) with (16), it is clear that removing  $\theta^{[j]}$  from (11) reduces the DMMM model to DCMLDA in (16). Moreover, if we further let

$$\theta_j \sim \text{Dir}(\mathbf{r}), \quad n_j \sim \text{NB}(r, p_j), \quad (17)$$

then the joint likelihood of  $\mathbf{x}_j$ ,  $\mathbf{z}_j$ , and  $n_j$  given  $\Phi$ ,  $\mathbf{r}$ , and  $p_j$  can be expressed as

$$P(\mathbf{x}_j, \mathbf{z}_j, n_j \mid \Phi, \mathbf{r}, p_j) = \frac{1}{n_j!} \prod_{v=1}^V \prod_{k=1}^K \frac{\Gamma(n_{vjk} + \phi_{vk} r_k)}{\Gamma(\phi_{vk} r_k)} p_j^{n_{vjk}} (1 - p_j)^{\phi_{vk} r_k}, \quad (18)$$

which multiplied by the combinatorial coefficient  $n_j! / (\prod_{v=1}^V \prod_{k=1}^K n_{vjk}!)$  is equal to the likelihood of  $\{n_{vj}, n_{vj1}, \dots, n_{vjK}\}_{v=1, V}$  given  $\Phi$ ,  $\mathbf{r}$ , and  $p_j$  in

Table 1: Predictive distributions for the multinomial mixed-membership model in (6), DCMLDA in (16), and Dirichlet-multinomial mixed-membership model in (11).

Model	Predictive distribution for $x_{ji}$	Predictive distribution for $z_{ji}$
Multinomial mixed-membership	$P(x_{ji} = v   \Phi, \theta_j) = \phi_{vz_{ji}}$	$P(z_{ji} = k   \theta_j) = \theta_{kj}/\theta_{\cdot j}$
DCMLDA	$P(x_{ji} = v   \mathbf{x}_j^{-ji}, z_{ji}, \Phi, \mathbf{r}) = \frac{n_{vjz_{ji}}^{-ji} + \phi_{vz_{ji}} r_{z_{ji}}}{n_{\cdot jz_{ji}}^{-ji} + r_{z_{ji}}}$	$P(z_{ji} = k   \theta_j) = \theta_{kj}/\theta_{\cdot j}$
Dirichlet-multinomial mixed-membership	$P(x_{ji} = v   \mathbf{x}_j^{-ji}, z_{ji}, \Phi, \theta_j) = \frac{n_{vjz_{ji}}^{-ji} + \phi_{vz_{ji}} \theta_{z_{ji}j}}{n_{\cdot jz_{ji}}^{-ji} + \theta_{z_{ji}j}}$	$P(z_{ji} = k   \mathbf{z}_j^{-i}, n_j, \theta_j) = \frac{n_{\cdot jk}^{-ji} + \theta_{kj}}{n_j - 1 + \theta_{\cdot j}}$

$$n_{vj} = \sum_{k=1}^K n_{vjk}, \quad n_{vjk} \sim \text{NB}(\phi_{vk} r_k, p_j). \quad (19)$$

Thus, as far as the conditional posteriors of  $\{\phi_k\}_k$  and  $\{r_k\}_k$  are concerned, DCMLDA constituted by (16)-(17) is equivalent to a special case of NBFA shown in (19), which is the augmented representation of  $\mathbf{n}_j \sim \text{NB}(\Phi \mathbf{r}, p_j)$  that restricts all instances to have the same factor scores  $\{r_k\}_k$  under the same set of shared factors  $\{\phi_k\}_k$ .

Given the factors  $\{\phi_k\}_k$  and factor scores  $\theta_j$ , for the multinomial mixed-membership model in (6), both the attribute and factor indices are independently drawn from the categorical distributions; for DCMLDA in (16), the factor indices but not the attribute indices are independently drawn from the categorical distributions; whereas for the DMMM model in (11), neither the factor indices nor attribute indices are independently drawn from the categorical distributions. Denoting  $y^{-ji}$  as the variable  $y$  obtained by excluding the contribution of the  $i$ th attribute token in instance  $j$ , we compare in Table 1 these three different models on their predictive distributions of  $x_{ji}$  and  $z_{ji}$ .

In comparison to the multinomial mixed-membership model, DCMLDA allows the number of times that an attribute appears in an instance to exhibit the “rich get richer” (*i.e.*, self-excitation) behavior, leading to a better modeling of attribute burstiness, and the DMMM model further models the burstiness of the factor indices and hence encourages not only

self-excitation, but also cross-excitation of attribute frequencies. It is clear from Table 1 that DCMLDA models attribute burstiness but not factor burstiness, and the corresponding NBFA restricts all instances to have the same factor scores under the shared factors  $\{\phi_k\}_k$ . Thus we expect the DMMM model to clearly outperform DCMLDA, as will be confirmed by our experimental results.

## 4 Hierarchical gamma-negative binomial process negative binomial factor analysis

Let  $G \sim \Gamma\text{P}(G_0, 1/c_0)$  be a gamma process (Ferguson, 1973; Kingman, 1993) defined on the product space  $\mathbb{R}_+ \times \Omega$ , where  $\mathbb{R}_+ = \{x : x > 0\}$ , with two parameters: a finite and continuous base measure  $G_0$  over a complete separable metric space  $\Omega$ , and a scale  $1/c_0$ , such that  $G(A) \sim \text{Gamma}(G_0(A), 1/c_0)$  for each  $A \subset \Omega$ . The Lévy measure of the gamma process can be expressed as  $\nu(dr d\phi) = r^{-1} e^{-c_0 r} dr G_0(d\phi)$ . A draw from  $G \sim \Gamma\text{P}(G_0, 1/c_0)$  can be represented as the countably infinite sum

$$G = \sum_{k=1}^{\infty} r_k \delta_{\phi_k}, \quad \phi_k \sim g_0,$$

where  $\gamma_0 = G_0(\Omega)$  as the mass parameter and  $g_0(d\phi) = G_0(d\phi)/\gamma_0$  is the base distribution.

To support countably infinite factors for the DMMM model, we consider a hierarchical gamma-negative binomial process (hGNBP) as

$$X_j \sim \text{NBP}(\Theta_j, p_j), \quad \Theta_j \sim \Gamma\text{P}(G, 1/c_j), \quad G \sim \Gamma\text{P}(G_0, 1/c_0), \quad (20)$$

where  $\Theta_j$  and  $X_j$  are the gamma and NB processes for instance  $j$ , respectively. Given a gamma process random draw  $G = \sum_{k=1}^{\infty} r_k \delta_{\phi_k}$ , we have

$$\Theta_j = \sum_{k=1}^{\infty} \theta_{kj} \delta_{\phi_k}, \quad \theta_{kj} \sim \text{Gamma}(r_k, 1/c_j), \quad (21)$$

where  $\theta_{kj} := \Theta_j(\phi_k)$  measures the weight of factor  $k$  in instance  $j$ , and

$$X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\phi_k}, \quad n_{jk} \sim \text{NB}(\theta_{kj}, p_j), \quad (22)$$

where  $n_j = X_j(\Omega)$  is the length of instance  $j$  and  $n_{jk} := X_j(\phi_k) = \sum_{i=1}^{n_j} \delta(z_{ji} = k)$  represents the number of times that factor  $k$  appears in instance  $j$ .

We provide the posterior analysis for the hGNBP in Appendix B.

## 4.1 Hierarchical model

We express the hGNBP-DMMM model as

$$\begin{aligned} x_{ji} &\sim \text{Cat}(\phi_{z_{ji}}^{[j]}), \quad z_{ji} \sim \text{Cat}(\theta^{[j]}), \\ \phi_k^{[j]} &\sim \text{Dir}(\phi_k \theta_{kj}), \quad \theta^{[j]} \sim \text{Dir}(\theta_j), \\ \theta_{kj} &\sim \text{Gamma}(r_k, 1/c_j), \quad c_j \sim \text{Gamma}(e_0, 1/f_0), \\ n_j &\sim \text{NB}(\theta_j, p_j), \quad p_j \sim \text{Beta}(a_0, b_0), \quad G \sim \Gamma\text{P}(G_0, 1/c_0), \end{aligned} \quad (23)$$

where the atoms of the gamma process are drawn from a Dirichlet base distribution  $\phi_k \sim \text{Dir}(\eta, \dots, \eta)$ . We further let  $\gamma_0 \sim \text{Gamma}(a_0, 1/b_0)$  and  $c_0 \sim \text{Gamma}(e_0, 1/f_0)$ .

With Proposition 2, the hGNBP-DMMM model in (23) can also be represented as a hGNBP-NBFA as

$$\begin{aligned} n_{vj} &= \sum_{k=1}^{\infty} n_{vjk}, \quad n_{vjk} \sim \text{NB}(\phi_{vk} \theta_{kj}, p_j), \\ \theta_{kj} &\sim \text{Gamma}(r_k, 1/c_j), \quad c_j \sim \text{Gamma}(e_0, 1/f_0), \\ p_j &\sim \text{Beta}(a_0, b_0), \quad G \sim \Gamma\text{P}(G_0, 1/c_0). \end{aligned} \quad (24)$$

The DMMM model in (23) and NBFA in (24) have the same conditional posteriors for both the factors  $\{\phi_k\}_k$  and factor scores  $\{\theta_j\}_j$ , but lead to different inference strategies. To infer  $\{\phi_k\}_k$  and  $\{\theta_j\}_j$ , we first develop both blocked and collapsed Gibbs sampling with (23), as

described in detail in Appendix C, and then develop blocked Gibbs sampling with (24), as described below. All three Gibbs sampling algorithms are summarized in Algorithm 1 shown in Appendix C.

#### 4.1.1 Blocked Gibbs sampling under compound Poisson representation

Examining both the blocked and collapsed Gibbs samplers presented Appendices C.1 and C.2, respectively, and their sampling steps shown in Algorithm 1, one may notice that to obtain  $n_{vjk}$  in each iteration, one has to go through all the attribute tokens  $x_{ji}$ , for each of which the sampling of  $z_{ji}$  from a multinomial distribution takes  $O(K)$  computation. However, as it is the  $\ell_{vjk}$  but not  $n_{vjk}$  that are required for posterior inference for all the other model parameters, one may naturally wonder whether the step of sampling  $z_{ji}$  to obtain  $n_{vjk}$  can be skipped. To answer that question, we first introduce the following theorem, whose proof is provided in Appendix A.

**Theorem 3.** *Conditioning on  $n$  and  $\mathbf{r}$ , with  $\{n_k\}_k$  marginalized out, the distribution of  $\boldsymbol{\ell} = (\ell_1, \dots, \ell_k)$  in*

$$\ell_k | n_k \sim \text{CRT}(n_k, r_k), \quad \mathbf{n} | n \sim \text{DirMult}(n, r_1, \dots, r_K)$$

*is the same as that in*

$$\boldsymbol{\ell} | \ell. \sim \text{Mult}(\ell., r_1/r., \dots, r_K/r.), \quad \ell. | n \sim \text{CRT}(n, r.),$$

*with PMF*

$$P(\boldsymbol{\ell} | n, r_1, \dots, r_K) = \frac{\ell.!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n + r.)} |s(n, \ell.)| \prod_{k=1}^K r_k^{\ell_k}.$$

Rather than representing NBFA in (24) as the DMMM model in (23), we may directly consider its compound Poisson representation as

$$n_{vj} = \sum_{t=1}^{\ell_{vj}} n_{vjt}, \quad n_{vjt} \sim \text{Log}(p_j), \quad \ell_{vj} \sim \text{Pois} \left[ - \sum_k \phi_{vk} \theta_{kj} \ln(1 - p_j) \right]. \quad (25)$$



Under this representation, we may first infer  $\ell_{vj}$  for each  $n_{vj}$  and then factorize the latent count matrix  $\{\ell_{vj}\}_{v,j}$  under the Poisson likelihood, as described below.

Rather than first sampling  $z_{ji}$  (and hence  $n_{vjk}$ ) using (C.1) and then sampling  $\ell_{vjk}$  using (C.2), as in Appendix C, with Theorem 3 and the compound Poisson representation in (25), we can skip sampling  $z_{ji}$  and directly sample  $\ell_{vjk}$  as

$$(\ell_{vj} | -) \sim \text{CRT} \left( n_{vj}, \sum_k \phi_{vk} \theta_{kj} \right), \quad (26)$$

$$[(\ell_{vj1}, \dots, \ell_{vjK}) | -] \sim \text{Mult} \left( \ell_{vj}, \frac{\phi_{v1} \theta_{1j}}{\sum_k \phi_{vk} \theta_{kj}}, \dots, \frac{\phi_{vK} \theta_{Kj}}{\sum_k \phi_{vk} \theta_{kj}} \right). \quad (27)$$

All the other model parameters can be sampled in the same way as they are sampled in the regular blocked Gibbs sampler, with (C.3)-(C.9) of Appendix C.

Note that  $\ell_{vj} = n_{vj}$  a.s. if  $n_{vj} \in \{0, 1\}$ ,  $\ell_{vj} \leq n_{vj}$  a.s. if  $n_{vj} \geq 2$ , and

$$\mathbb{E}[\ell_{vj} | n_{vj}, \phi_{vk} \theta_{kj}] = (\sum_k \phi_{vk} \theta_{kj}) [\psi(n_{vj} + \sum_k \phi_{vk} \theta_{kj}) - \psi(\sum_k \phi_{vk} \theta_{kj})],$$

where  $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$  is the digamma function. Thus  $\mathbb{E}[\ell_{vj}] \approx (\sum_k \phi_{vk} \theta_{kj}) \ln(n_{vj})$  when  $n_{vj}$  is large. Clearly, this new sampling strategy significantly impacts  $n_{vj}$  that are large. In comparison to the regular blocked Gibbs sampler described in detail in Appendix C, the compound Poisson based blocked Gibbs sampler essentially replaces (C.1)-(C.2) of the regular blocked Gibbs sampler with (26)-(27), which can be readily justified by Theorem 3. Instead of directly assigning the  $n_{vj}$  attribute tokens to the  $K$  latent factors, now we only need to assign them to  $O[\ln(n_{vj} + 1)]$  tables and directly assign these tables to latent factors. As assigning an attribute token to a table can be accomplished by just a single Bernoulli random draw, this new sampling procedure reduces the computational complexity for sampling all  $\ell_{vjk}$  from  $O(n..K)$  to  $O[\sum_v \sum_j \ln(n_{vj} + 1)K]$ , which could lead to a considerable saving in computation for long instances where large counts  $n_{vj}$  are abundant. This new sampling algorithm not only is less expensive in computation, but also may converge much faster as there is no need to worry about the dependencies between the MCMC samples for the factor

Table 2: Comparisons of the GNB-PFA, GNB-DCMLDA, and hGNBP-NBFA.

	GNBP-PFA (multinomial mixed-membership model)	GNBP-DCMLDA	hGNBP-NBFA (Dirichlet-multinomial mixed-membership model)
Estimated Poisson rate of $n_{vj}$ given the factors and factor scores	$\sum_k \phi_{vk} \theta_{kj}$	$(n_{vj} + \sum_k \phi_{vk} r_k) p_j$	$(n_{vj} + \sum_k \phi_{vk} \theta_{kj}) p_j$
Factor analysis	$\mathbf{n}_j \sim \text{Pois}(\mathbf{\Phi} \boldsymbol{\theta}_j)$	$\mathbf{n}_j \sim \text{NB}(\mathbf{\Phi} \mathbf{r}, p_j)$	$\mathbf{n}_j \sim \text{NB}(\mathbf{\Phi} \boldsymbol{\theta}_j, p_j)$
Mixed-membership modeling	$x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}),$ $z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j / \theta_{.j})$	$x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}),$ $z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j),$ $\boldsymbol{\phi}_k^{[j]} \sim \text{Dir}(\boldsymbol{\phi}_k r_k),$ $\boldsymbol{\theta}_j \sim \text{Dir}(\mathbf{r})$	$x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}),$ $z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j^{[j]}),$ $\boldsymbol{\phi}_k^{[j]} \sim \text{Dir}(\boldsymbol{\phi}_k \theta_{kj}),$ $\boldsymbol{\theta}_j^{[j]} \sim \text{Dir}(\boldsymbol{\theta}_j)$
Distribution of $X_j = \sum_{k=1}^{\infty} n_{.jk} \delta_{\phi_k}$ given $G$	$X_j   G, p_j \sim \text{NBP}(G, p_j)$	$X_j   G, p_j \sim \text{NBP}(G, p_j)$	$X_j   G, c_j, p_j \sim \text{GNBP}(G, c_j, p_j)$
Variance-to-mean ratio of $n_{.jk}$ given $c_j$ and $p_j$	$\frac{1}{1 - p_j}$	$\frac{1}{1 - p_j}$	$\frac{1}{1 - p_j} + \frac{p_j}{c_j(1 - p_j)^2}$

indices  $z_{ji}$ , which are not used at all under the compound Poisson representation.

## 4.2 Model comparison

We describe in detail in Appendix D that the gamma-negative binomial process (GNBP) can be used as a nonparametric Bayesian prior for both PFA and DCMLDA. In the prior, for PFA, we have  $n_{vj} \sim \text{Pois}(\sum_k \phi_{vk} \theta_{kj})$ , whereas for NBFA, we have  $n_{vj} \sim \text{NB}(\sum_k \phi_{vk} \theta_{kj}, p_j)$ , which can be augmented as

$$n_{vj} \sim \text{Pois}(\lambda_{vj}), \quad \lambda_{vj} \sim \text{Gamma}[\sum_k \phi_{vk} \theta_{kj}, p_j / (1 - p_j)].$$

Thus we have  $(\lambda_{vj} | n_{vj}, \mathbf{\Phi}, \boldsymbol{\theta}_j, p_j) \sim \text{Gamma}(n_{vj} + \sum_k \phi_{vk} \theta_{kj}, p_j)$  for NBFA. Similarly, we have  $(\lambda_{vj} | -) \sim \text{Gamma}(n_{vj} + \sum_k \phi_{vk} r_k, p_j)$  for the GNB-DCMLDA. To better understand the similarities and differences between the GNB-PFA, GNB-DCMLDA and hGNBP-NBFA, in Table 2 we compare their Poisson rates of  $n_{vj}$ , estimated with the factors and factor scores in a single MCMC sample, and other important model properties.

To estimate the latent Poisson rates for each count  $n_{vj}$  and hence the smoothed normalized attribute frequencies, it is clear from the second row of Table 2 that PFA (the multinomial mixed-membership model) solely relies on the inferred factors  $\{\boldsymbol{\phi}_k\}$  and fac-

tor scores  $\{\theta_{kj}\}$ , DCMLDA adds an instance-invariant smoothing parameter, calculated as  $\sum_v \phi_{vk} r_k$ , into the observed count  $n_{vj}$  and weights that sum by an instance-specific probability parameter  $p_j$ , whereas NBFA (the DMMM model) adds an instance-specific smoothing parameter, calculated as  $\sum_v \phi_{vk} \theta_{kj}$ , into the observed count  $n_{vj}$  and weights that sum by  $p_j$ . Thus PFA represents an extreme that the observed counts are used to infer the factors and factor scores but are not used to directly estimate the Poisson rates; DCMLDA represents another extreme that the attribute frequencies in all instances are indiscriminately smoothed by the same set of smoothing parameters; whereas NBFA combines the observed counts with the inferred instance-specific smoothing parameters.

## 5 Example Results

We apply the proposed models to factorize attribute-instance count matrices, each column of which is represented as a  $V$  dimensional attribute-frequency count vector, where  $V$  is the number of unique attributes. We set the hyper-parameters as  $a_0 = b_0 = 0.01$  and  $e_0 = f_0 = 1$ . We consider the JACM<sup>1</sup>, Psychological Review<sup>2</sup> (PsyReview), and NIPS12<sup>3</sup> datasets, choosing attributes that occur in five or more instances. In addition, we consider the 20newsgroups dataset<sup>4</sup>, consisting of 18,774 instances from 20 different categories. It is partitioned into a training set of 11,269 instances and a testing set of 7,505 ones that were collected at later times. We remove a standard list of stopwords and attributes that appear less than five times. As summarized in Table 3, for the PysReview and JACM datasets, each of whose instance corresponds to the abstract of a research paper, the average instance lengths are only about 56 and 127, respectively. By contrast, a NIPS12 instance that includes the words of all sections of a research paper is in average more than ten times longer. By varying the percentage of attribute tokens randomly selected from each instance for training, we construct a set of attribute-instance matrices with a large variation on the

---

<sup>1</sup><http://www.cs.princeton.edu/~blei/downloads/>

<sup>2</sup>[http://psiexp.ss.uci.edu/research/programs\\_data/toolbox.htm](http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm)

<sup>3</sup><http://www.cs.nyu.edu/~roweis/data.html>

<sup>4</sup><http://qwone.com/~jason/20Newsgroups/>

average lengths of instances, which will be used to help make comparison between different models. Depending on applications, we either treat the Dirichlet smoothing parameter  $\eta$  as a tuning parameter, or sample it via data augmentation, as described in Appendix E.

To learn the factors in all the following experiments, we use the compound Poisson representation based blocked Gibbs sampler for both the hGNBP-NBFA and GNBP-NBFA, and use collapsed Gibbs sampling for the GNBP-PFA. We compare different samplers for the hGNBP-NBFA and provide justifications for choosing these samplers in Appendix F.

Table 3: Datasets used in experiments.

	JACM	PsyReview	NIPS12	20newsgroups
Number of unique attributes $V$	1,539	2,566	13,649	33,420
Number of instances	536	1,281	1,740	11,269
Total number of attribute tokens	68,055	71,279	2,301,375	2,351,800
Average instance length	127	56	1,323	209

## 5.1 Prediction of heldout attribute tokens

### 5.1.1 Experimental settings

We randomly choose a certain percentage of the attribute tokens in each instance as training, and use the remaining ones to calculate heldout perplexity. As shown in Zhou and Carin (2015), the GNBP-PFA performs similarly to the hierarchical Dirichlet process LDA of Teh et al. (2006) and outperforms a wide array of discrete latent variable models, thus we choose it for comparison. To demonstrate the importance of modeling both the burstiness of the attributes and that of the factors, we also make comparison to the GNBP-DCMLDA that considers only attribute burstiness. Since the inferred number of factors and hence the performance often depends on the Dirichlet smoothing parameter  $\eta$ , we set  $\eta$  as 0.005, 0.02, 0.01, 0.05, 0.10, 0.25, or 0.50. We vary both the training percentage and  $\eta$  to examine how the average instance length and the value of  $\eta$  influence the behaviors of the GNBP-PFA, GNBP-DCMLDA, and hGNBP-NBFA and impact their performance relative to each other.

For all three algorithms, we initialize the number of factors as  $K = 400$  and consider 5000 Gibbs sampling iterations, with the first 2500 samples discarded and every sample

per five iterations collected afterwards. For each collected sample, for the GNB-PFA, we draw the factors  $(\phi_k | -) \sim \text{Dir}(\eta + n_{1.k}, \dots, \eta + n_{V.k})$  and factor scores  $(\theta_{kj} | -) \sim \text{Gamma}(n_{jk} + r_k, p_j)$  for  $k \in \{1, \dots, K^+ + K_\star\}$ , where we let  $n_{v.k} = 0$  for all  $k > K^+$ ; for the GNB-DCMLDA, we draw the factors  $(\phi_k | -) \sim \text{Dir}(\eta + \ell_{1.k}, \dots, \eta + \ell_{V.k})$  and the weights  $(r_k | -) \sim \text{Gamma}(\ell_{..k}, 1/[c_0 - \sum_j \ln(1 - p_j)])$ , where we let  $\ell_{v.k} = 0$  and  $\ell_{..k} = \gamma_0/K_\star$  for all  $k > K^+$ ; and for the hGNBP-NBFA, we draw the factors  $(\phi_k | -) \sim \text{Dir}(\eta + \ell_{1.k}, \dots, \eta + \ell_{V.k})$  and factor scores  $(\theta_{kj} | -) \sim \text{Gamma}[r_k + \ell_{.jk}, 1/(c_j - \ln(1 - p_j))]$ , where we let  $\ell_{v.k} = 0$  and  $r_k = \gamma_0/K_\star$  for all  $k > K^+$ . We set  $K_\star = 20$  for all three algorithms.

We compute the heldout perplexity as

$$\exp \left( -\frac{1}{m_{..}^{\text{test}}} \sum_v \sum_j m_{vj}^{\text{test}} \ln \frac{\sum_s \lambda_{vj}^{(s)}}{\sum_s \sum_v \lambda_{vj}^{(s)}} \right),$$

where  $s \in \{1, \dots, S\}$  is the index of a collected MCMC sample,  $m_{vj}^{\text{test}}$  is the number of test attribute tokens at attribute  $v$  in instance  $j$ ,  $m_{..}^{\text{test}} = \sum_v \sum_j m_{vj}^{\text{test}}$ , and  $\lambda_{vj}^{(s)}$  are computed using the equations shown in the second row of Tabel 2, *e.g.*, we have  $\lambda_{vj}^{(s)} = \left( n_{vj} + \sum_{k=1}^{K^+ + K_\star} \phi_{vk}^{(s)} \theta_{kj}^{(s)} \right) p_j^{(s)}$  for the hGNBP-NBFA. For each unique combination of  $\eta$  and the training percentage, the results are averaged over five random training/testing partitions. The evaluation method is similar to those used in Newman et al. (2009); Wallach et al. (2009); Paisley et al. (2012), and Zhou et al. (2012). All algorithms are coded in MATLAB, with the steps of sampling factor and table indices coded in C to optimize speed. We terminate a trial and omit the results for that particular setting if it takes a single core of an Intel Xeon 3.3 GHz CPU more than 24 hours to finish 5000 iterations. The code will be made available in the author's website for reproducible research.

We first consider the NIPS12 dataset, whose average instance length is about 1323, and present its results in Figures 1-4. We also consider both the PsyReview and JACM datasets, whose average instance lengths are about 56 and 127, respectively, and provide related plots in Appendix G.

### 5.1.2 General observations

For multinomial mixed-membership models, generally speaking, the smaller the Dirichlet smoothing parameter  $\eta$  is, the more sparse and specific the inferred factors are encouraged to be, and the larger the number of inferred factors using a nonparametric Bayesian mixed-membership modeling prior, such as the hierarchical Dirichlet process and the gamma- and beta-negative binomial processes (Paisley et al., 2012; Zhou et al., 2012). As shown in Figures 1(a)-(e), for the hGNBP-NBFA, a nonparametric Bayesian DMMM model, we observe a relationship between the number of inferred factors and  $\eta$  similar to that for the GNBP-PFA, a nonparametric Bayesian multinomial mixed-membership model. In comparison to multinomial mixed-membership models such as the GNBP-PFA, what make the hGNBP-NBFA different and desirable are:

- 1) its parsimonious representation that uses fewer factors to achieve better heldout prediction, as shown in Figures 2(a)-(e);
- 2) its distinct mechanism in adjusting its number of inferred factors according to the lengths of instances, as shown in Figures 3(a)-(f);
- 3) its significantly lower computational complexity for an attribute-instance matrix with long instances (large column sums), with the differences becoming increasingly more significant as the the average instance length increases, as shown in Figures 2(f)-(j) and 4(a)-(f);
- 4) its ability to achieve the same predictive power with significantly less time, as shown in Figures 4(g)-(l);
- 5) and its overall better predictive performance both under various values of  $\eta$  while controlling the instance lengths, as shown in Figures 1(f)-(j), and under various instance lengths while controlling  $\eta$ , as shown in Figures 3(g)-(l).

### 5.1.3 Detailed discussions

**Distinct behavior and parsimonious representation.** When fixing  $\eta$  but gradually increasing the average instance length, the number of factors inferred by a nonparametric

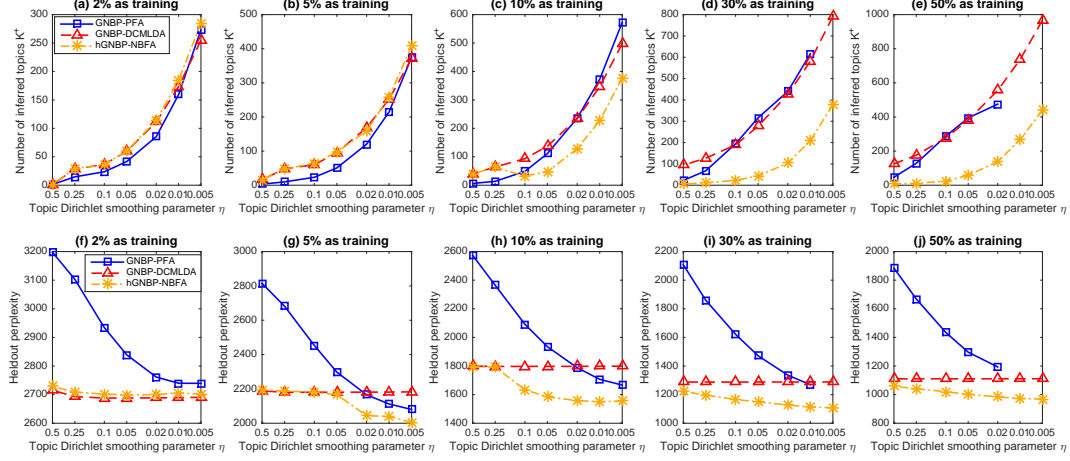


Figure 1: Comparisons of the GNPB-PFA (multinomial mixed-membership model), GNPB-DCMLDA, and hGNBP-NBFA (Dirichlet-multinomial mixed-membership model) on (a)-(e) the posterior means of the number of inferred factors  $K^+$  and (f)-(j) heldout perplexity, both as a function of the Dirichlet smoothing parameter  $\eta$  for the NIPS12 dataset. The values of  $\eta$  are plot in the logarithmic scale from large to small. In both rows, the plots from left to right are obtained using 2%, 5%, 10%, 30%, and 50% of the attribute tokens for training, respectively. All plots are based on five independent random trials. The error bars are not shown as variations across different trials are small. Some results for the GNPB-PFA are missing as they took more than 24 hours to run 5000 Gibbs sampling iterations on a 3.3 GHz CPU and hence were terminated before completion.

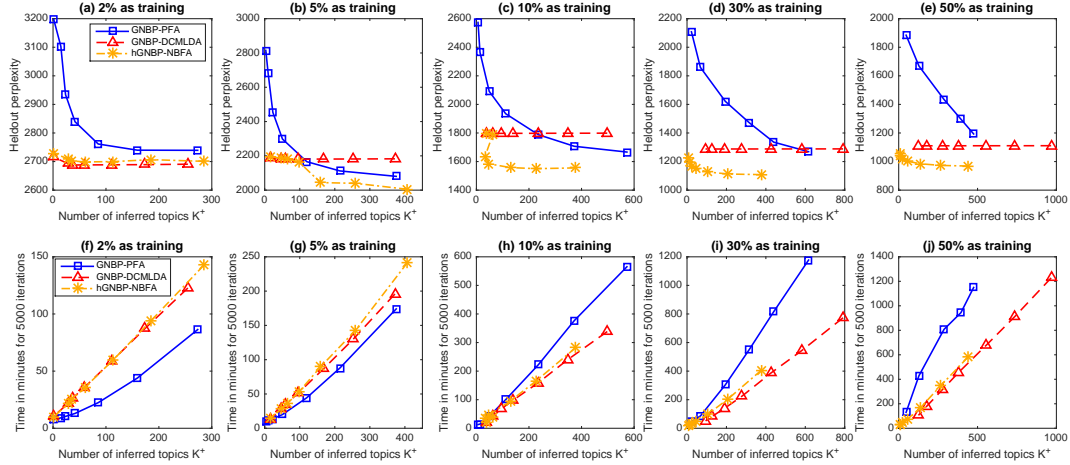


Figure 2: Using the same results shown in Figure 1, we plot (a)-(e) the obtained heldout perplexity and (f)-(j) the number of minutes to finish 5000 Gibbs sampling iterations, both as a function of the number of inferred factors  $K^+$ .

Bayesian multinomial mixed-membership model such as the GNPB-PFA often increases at a near-constant rate, as shown with the blue curves in Figure 3(a)-(f). The GNPB-DCMLDA, which models attribute burtiness, behaves similarly in the number of inferred factors, as

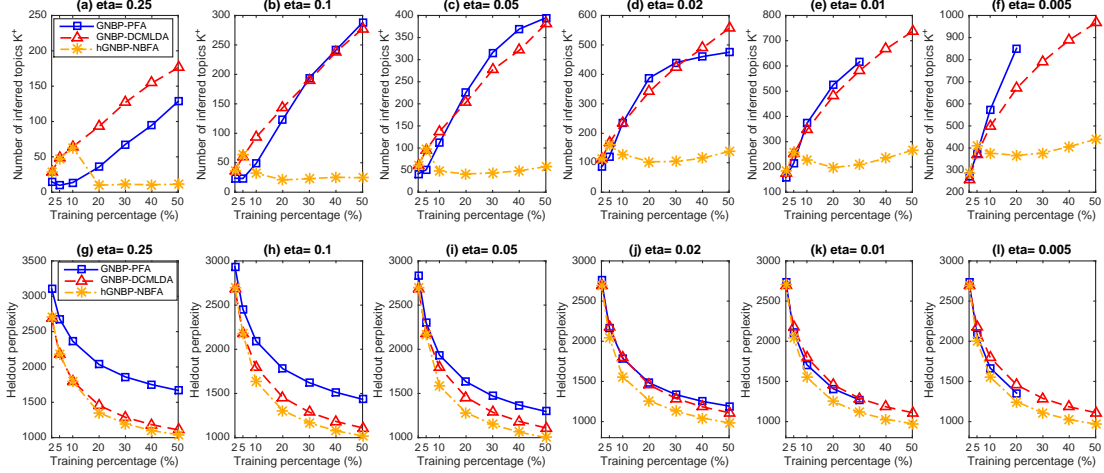


Figure 3: Comparisons of the GNPB-PFA (multinomial mixed-membership model), GNPB-DCMLDA, and hGNBP-NBFA (Dirichlet-multinomial mixed-membership model) on (a)-(f) the posterior means of the number of inferred factors  $K^+$  and (g)-(l) heldout perplexity, both as a function of the percentage of attribute tokens used for training for the NIPS12 dataset. In both rows, the plots from left to right are obtained with  $\eta = 0.25, 0.1, 0.05, 0.02, 0.01$ , and  $0.005$ , respectively. Other specifications are the same as those of Figure 1.

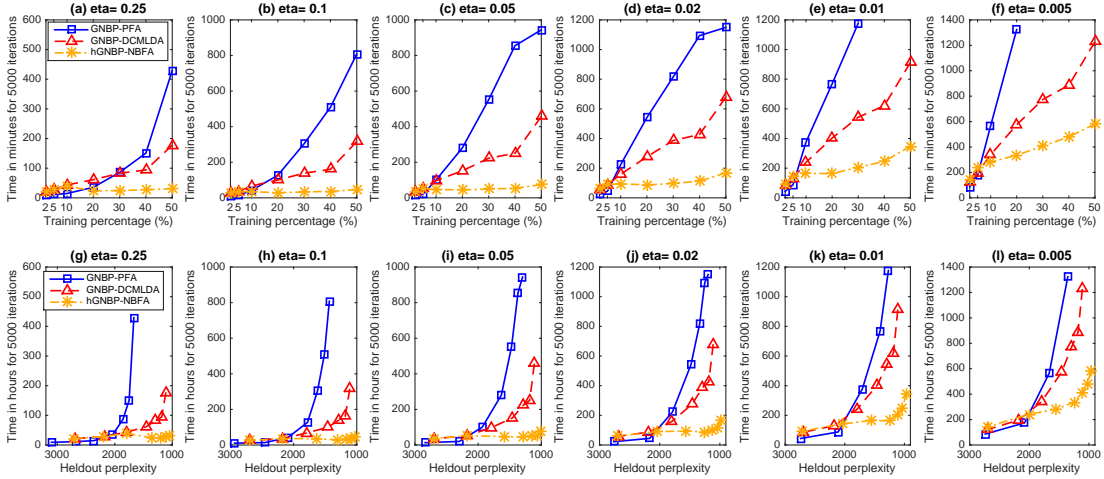


Figure 4: Using the results shown in Figure 3, we plot the number of minutes to finish 5000 Gibbs sampling iterations both (a)-(f) as a function of the percentage of attributes used for training and (g)-(l) as a function of heldout perplexity.

shown with the red curves in Figure 3(a)-(f). Under the same setting, the number of inferred factors by the hGNBP-NBFA often first increases at a similar near-constant rate when the average instance length is short, however, it starts decreasing once the average instance length becomes sufficiently long, and eventually turns around and increases, but at a much lower rate, as the average instance length further increases, as shown with the yellow curves



in Figure 3(a)-(f). This distinct behavior implies that by exploiting its ability to model both attribute and factor burstiness, the hGNBP-NBFA could have a parsimonious representation of a dataset with long instances. By contrast, a nonparametric Bayesian multinomial mixed-membership model, such as the GNB-PFA, models neither attribute nor factor burstiness. Consequently, it has to increase its latent dimension at a near-constant rate as a function of the average instance length, in order to adequately capture self- and cross-excitation of attribute frequencies, which are often more prominent in longer instances. It is clear that by decreasing  $\eta$  and hence increasing the number of inferred factors, the GNB-PFA can gradually approach and eventually outperform the GNB-DCMLDA, but still clearly underperform the hGNBP-NBFA in most cases, even if using many more factors and consequently significantly more computation.

**Combining factorization and the modeling of burstiness.** As shown in Figure 1(f), when the training percentage is as small as 2%, the GNB-DCMLDA, which combines the observed counts  $n_{vj}$  and the inferred instance-invariant smoothing parameters  $\sum_k \phi_{vk} r_k$  to estimate the Poisson rates (and hence the smoothed normalized attribute frequencies), achieves the best predictive performance (lowest heldout perplexity); the hGNBP-NBFA tries to improve DCMLDA by combining the observed counts and document-specific smoothing parameters  $\sum_k \phi_{vk} \theta_{kj}$ , and the GNB-PFA only relies on  $\sum_k \phi_{vk} \theta_{kj}$ , yielding slightly and significantly worse performance, respectively, at this relatively extreme setting. This suggests that when the observed counts are too small, using factorization may not provide any advantages than simply smoothing the raw attribute counts with instance-invariant smoothing parameters.

As the training percentage increases, all three algorithms quickly improve their performance, as shown in Figures 1(g)-(j). Given a training percentage that is sufficiently large, e.g., 10% for this dataset (*i.e.*, the average training instance length is about 132), all three algorithms tend to increase their numbers of inferred factors  $K^+$  as  $\eta$  decreases, although the hGNBP-NBFA usually has a lower increasing rate. They differ from other each signifi-

cantly, however, on how the performance improves as the inferred number factors increases, as shown in Figures 2(a)-(e): for the GNB-DCMLDA, as it relies on  $\sum_k \phi_{vk} r_k$  to smooth the observed counts, its predictive power is almost invariant to the change of  $\eta$  and its number of factors; for the GNB-PFA, by decreasing  $\eta$  and hence increasing its number of inferred factors, it can approach and eventually outperform DCMLDA; whereas for the hGNB-NBFA, it follows DCMLDA closely when  $\eta$  is large or the lengths of instances are short, but often reduces its rate of increase for the number of inferred factors as  $\eta$  decreases and quickly lowers its perplexity as  $K^+$  increases, as long as  $\eta$  is sufficiently small or the instances are sufficiently long. Thus in general, the hGNB-NBFA provides the lowest perplexity using the least number of inferred factors.

Note that when the lengths of the training instances are short, setting  $\eta$  to be large will make the factors  $\phi_k$  of NBFA become over-smoothed and hence NBFA becomes essentially the same as DCMLDA. As  $\eta$  decreases given the same average instance length, or as the average instance length increases given the same  $\eta$ , the factorization of NBFA with instance-dependent factor scores gradually take effect to improve the estimation of the Poisson rates and hence the smoothed normalized attribute frequencies for each instance. Overall, by combining the factorization, as used in PFA, the modeling of attribute burstiness, as used in DCMLDA, and the modeling of factor burstiness, unique to NBFA, the hGNB-NBFA captures both self- and cross-excitation of attribute frequencies and achieves the best predictive performance with the most parsimonious representation as long as the average instance length is not too short and the value of  $\eta$  is not set too large to overly smooth the factors.

**Significantly lower computation for sufficiently long instances.** For the GNB-PFA, the collapsed Gibbs sampler samples all the factor indices with a computational complexity of  $O(n..K^+)$ , whereas for the hGNB-NBFA, the corresponding computation has a complexity of  $O[\sum_v \sum_j \ln(n_{vj} + 1)K^+]$  and sampling  $\{\phi_k\}_k$  and  $\{\theta_j\}_j$  adds an additional computation of  $O(VK^+ + NK^+)$ . Thus the computation for the hGNB-NBFA not only is often lower given the same  $K^+$  for a dataset consisting of sufficiently long instances,

but also becomes much lower because the inferred  $K^+$  is often much smaller when the instance lengths are sufficiently long. For example, as shown in Figure 2(i), when 30% of the attribute tokens in each instance are used for training, which means the average training instance length is about 397, the time for the GNB-PFA to finish 5000 Gibbs sampling iteration on a 3.3 GHz CPU is about double that for the hGNBP-NBFA when their inferred numbers of factors are similar to each other; and when 20% of the attribute tokens in each instance are used for training (*i.e.*, the average training instance length is around 265), in comparison to the hGNBP-NBFA, the GNB-PFA takes about three times more minutes when  $\eta = 0.1$ , as shown in Figures 4(b), and four times more minutes when  $\eta = 0.01$ , as shown in Figures 4(e). Overall, for a dataset whose instances are not too short to exhibit self- and cross-excitation of attribute frequencies, the hGNBP-NBFA often takes the least time to finish the computation while controlling the value of  $\eta$  and average instance length, has lower computation given the same inferred number of factors  $K^+$ , and achieves a low perplexity with significantly less computation.

## 5.2 Unsupervised feature learning for classification

To further verify the advantages of NBFA that models both self- and cross-excitation of attribute frequencies, we use the proposed models to extract low-dimensional feature vectors from high-dimensional attribute-frequency count vectors of the 20newsgroups dataset, and then examine how well the unsupervisedly extracted feature vector of a test instance can be used to correctly classify it to one of the 20 news groups. As the classification accuracy often strongly depends on the dimension of the feature vectors, we truncate the total number of factors at  $K = 25, 50, 100, 200, 400, 600, 800$ , or 1000. Correspondingly, we slightly modify the gamma process based nonparametric Bayesian models by choosing a discrete base measure for the gamma process as  $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K} \delta_{\phi_k}$ ,  $\phi_k \sim \text{Dir}(\eta, \dots, \eta)$ . Thus in the prior we now have  $r_k = G(\phi_k) \sim \text{Gamma}(\gamma_0/K, 1/c_0)$  and consequently the Gibbs sampling update equations for  $\{r_k\}_k$  and  $\gamma_0$  will also slightly change. We omit these details for brevity, and refer to Zhou and Carin (2015) on how the same type of finite truncation is

used in inference for nonparametric Bayesian models.

For this application, we fix the truncation level  $K$  but impose the  $\text{Gamma}(0.01, 1/0.01)$  prior on the Dirichlet smoothing parameter  $\eta$ , letting it be inferred from the data using (E.4). The same as before, we consider collapsed Gibbs sampling for the GNB-PFA and the compound Poisson representation based blocked Gibbs sampler for the hGNBP-NBFA, with the main difference in that a fixed instead of an adaptive truncation is now used for inference. We do not consider the GNB-DCMLDA here since it does not provide instance specific feature vectors under the same set of shared factors. Note that although we fix  $K$ , if  $K$  is set to be large enough, not necessarily all factors would be used and hence a truncated model still preserves its ability to infer the number of active factors  $K^+ \leq K$ ; whereas if  $K$  is set to be small, a truncated model may lose its ability to infer  $K^+$ , but it still maintains asymmetric priors (Wallach et al., 2009) on the factor scores.

For both the hGNBP-NBFA and GNB-PFA, we consider 2000 Gibbs sampling iterations on the 11,269 training instances of the 20newsgroups dataset, and retain the weights  $\{r_k\}_{1,K}$  and the posterior means of  $\{\phi_k\}_{1,K}$  as factors, according to the last MCMC sample, for testing. With these  $K$  inferred factors and weights, we further apply 1000 blocked Gibbs sampling iterations for both models and collect the last 500 MCMC samples to estimate the posterior mean of the feature usage proportion vector  $\theta_j/\theta_{\cdot,j}$ , for every instance in both the training and testing sets. Denote  $\bar{\theta}_j \in \mathbb{R}^K$  as the estimated feature vector for instance  $j$ . We use the  $L_2$  regularized logistic regression provided by the LIBLINEAR<sup>5</sup> package (Fan et al., 2008) to train a linear classifier on all  $\bar{\theta}_j$  in the training set and use it to classify each  $\bar{\theta}_j$  in the test set to one of the 20 news groups; the regularization parameter  $C$  of the classifier five-fold cross validated on the training set from  $(2^{-10}, 2^{-9}, \dots, 2^{15})$ .

We first consider distinguishing between the *alt.atheism* and *talk.religion.misc* news groups, and between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* news groups. For each binary classification task, we remove a standard list of stop words and only consider

---

<sup>5</sup><https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

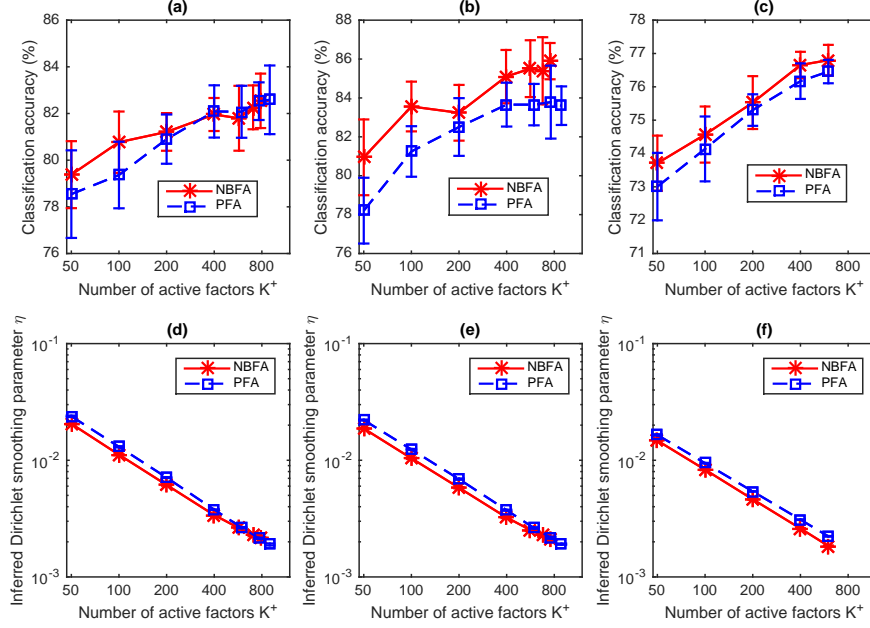


Figure 5: Comparison between negative binomial factor analysis (NBFA) and Poisson factor analysis (PFA) on three different classification tasks, with the number of factors  $K$  fixed and the Dirichlet smoothing parameter  $\eta$  inferred from the data. For the *alt.atheism* versus *talk.religion.misc* binary classification task, based on twelve independent random trials, we plot (a) the classification accuracy and (d) the inferred  $\eta$ , both as a function of the number active topics  $K^+ \leq K$ , where  $K \in \{50, 100, 200, 400, 600, 800, 1000\}$ . (d) and (e): Analogous plots to (a) and (b) for the *comp.sys.ibm.pc.hardware* versus *comp.sys.mac.hardware* binary classification task, with  $K \in \{50, 100, 200, 400, 600, 800, 1000\}$ . (c) and (f): Analogous plots to (a) and (b) for the 20newsgroups multi-class classification task, with  $K \in \{50, 100, 200, 400, 600\}$ .

the attributes that appear at least five times in both newsgroups combined, and report the classification accuracies based on twelve independent runs with random initializations.

As shown in Figures 5(a) and 5(b), NBFA clearly outperforms PFA for both binary classification tasks in that it in general provides higher classification accuracies on testing instances while controlling the truncation level  $K$  (*i.e.*, the dimension of the extracted feature vectors). It is also interesting to examine how the inferred Dirichlet smoothing parameter  $\eta$  changes as the truncation level  $K$  increases, as shown in Figures 5(d) and 5(e). It appears that the inferred  $\eta$ 's and active factors  $K^+$ 's could be fitted with a decreasing straight line in the logarithmic scale, except for the tails that seem slightly concave up, for both NBFA and PFA. When the truncation level  $K$  is not sufficiently large, the inferred  $\eta$  of NBFA is usually smaller than that of PFA given the same  $K$ . This may be explained by examining

(E.4), where  $\ell_{vjk} \leq n_{vjk}$  a.s. and the differences could be significant for large  $n_{vjk}$ .

In addition to these two binary classification tasks, we consider multi-class classification on the 20newsgroups dataset. After removing stopwords and attributes that appear less than five times, we obtain 33,420 unique attributes and over 2 million attribute tokens, as summarized in Table 3. We use all 11,269 training instances to infer the factors and factor scores, and mimic the same testing procedure used for binary classification to extract low-dimensional feature vectors, with which each testing instance is classified to one of the 20 news groups using the same  $L_2$  regularized logistic regression. As shown in Figure 5(f), NBFA generally outperforms PFA in terms of classification accuracies given the same feature dimensions, consistent with our observations for both binary classification tasks. We also observe similar relationship between the  $K^+$  and inferred  $\eta$  as we do in both binary classification tasks.

## 6 Conclusions

Negative binomial factor analysis (NBFA) is proposed to factorize the attribute-instance count matrix under the NB likelihood. Its equivalent representation as the Dirichlet multinomial mixed-membership model reveals how the modeling of not only the burstiness of attributes, but also that of the factors distinguishes it from previously proposed discrete latent variable models. The hierarchical gamma-negative binomial process (hGNBP) is further proposed to support NBFA with countably infinite factors, and a compound Poisson representation based blocked Gibbs sampler, which adaptively truncates the number of factors in each MCMC iteration, is shown to converge fast and have low computational complexity. By capturing both self- and cross-excitation of attribute frequencies and by smoothing the observed counts with both instance and attribute specific rates obtained through factorization under the NB likelihood, the hGNBP-NBFA not only infers a parsimonious representation of an attribute-instance count matrix, but also achieves state-of-the-art predictive performance at low computational cost. In addition, the latent feature vectors inferred under the

hGNBP-NBFA are better suited for classification than those inferred by the GNPB-PFA. It is of interest to investigate a wide variety of extensions built on Poisson factor analysis under this new modeling framework.

## References

- Acharya, A., J. Ghosh, and M. Zhou (2015). Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*, pp. 1–9.
- Adams, R. P., Z. Ghahramani, and M. I. Jordan (2010). Tree-structured stick breaking for hierarchical data. In *NIPS*.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* 2(6), 1152–1174.
- Blei, D. and J. Lafferty (2006a). Correlated topic models. *NIPS*.
- Blei, D., A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.
- Blei, D. M., T. L. Griffiths, and M. I. Jordan (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of ACM*.
- Blei, D. M. and J. D. Lafferty (2006b). Dynamic topic models. In *ICML*.
- Broderick, T., L. Mackey, J. Paisley, and M. I. Jordan (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Buntine, W. and A. Jakulin (2006). Discrete component analysis. In *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag.
- Canny, J. (2004). Gap: a factor model for discrete data. In *SIGIR*.
- Caron, F., Y. W. Teh, and B. T. Murphy (2014). Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data. *Annals of Applied Statistics*.
- Church, K. W. and W. A. Gale (1995). Poisson mixtures. *Natural Language Engineering*.
- Doyle, G. and C. Elkan (2009). Accounting for burstiness in topic models. In *ICML*.
- Dunson, D. B. and A. H. Herring (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* 6(1), 11–25.
- Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*
- Ewens, W. J. (1972). *Theoretical population biology* 3(1), 87–112.
- Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). LIBLINEAR: A library for large linear classification. *JMLR*, 1871–1874.
- Favaro, S. and Y. W. Teh (2013). MCMC for normalized random measure mixture models. *Statistical Science* 28(3), 335–359.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1(2), 209–230.

- Fisher, R. A., A. S. Corbet, and C. B. Williams (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12(1), 42–58.
- Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*.
- Gan, Z., C. Chen, R. Henao, D. Carlson, and L. Carin (2015). Scalable deep poisson factor analysis for topic modeling. In *ICML*.
- Griffin, J. E. and S. G. Walker (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics* 20(1), 241–259.
- Griffiths, T. L. and Z. Ghahramani (2005). Infinite latent feature models and the Indian buffet process. In *NIPS*.
- Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *PNAS*.
- Heaukulani, C. and D. M. Roy (2015). The combinatorial structure of beta negative binomial processes. *to appear in Bernoulli*.
- Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *UAI*.
- Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* 96(453).
- James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*.
- James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* 36(1), 76–97.
- Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete multivariate distributions*, Volume 165. Wiley New York.
- Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press.
- Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. In *NIPS*.
- Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B* 69(4), 715–740.
- Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker (Eds.), *Bayesian nonparametrics*. Cambridge University Press.
- Lo, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Zeitschrift fur*, 55–66.
- Madsen, R. E., D. Kauchak, and C. Elkan (2005). Modeling word burstiness using the Dirichlet distribution. In *ICML*.
- Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate  $\beta$ -distribution, and correlations among proportions. *Biometrika*, 65–82.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*.



- Newman, D., A. Asuncion, P. Smyth, and M. Welling (2009). Distributed algorithms for topic models. *JMLR*.
- Paisley, J., C. Wang, and D. M. Blei (2012). The discrete infinite logistic normal distribution. *Bayesian Analysis*.
- Paisley, J., C. Wang, D. M. Blei, and M. I. Jordan (2015). Nested hierarchical dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*.
- Pitman, J. (2006). *Combinatorial stochastic processes*. Lecture Notes in Mathematics. Springer-Verlag.
- Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2), 945–959.
- Ranganath, R. and D. Blei (2015). Correlated random measures. *arXiv:1507.00720v1*.
- Ranganath, R., L. Tang, L. Charlin, and D. M. Blei (2015). Deep exponential families. In *AISTATS*.
- Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* 31(2), 560–585.
- Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*
- Titsias, M. K. (2008). The infinite gamma-Poisson feature model. In *NIPS*.
- Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics Simulation and Computation*.
- Wallach, H. M., D. M. Mimno, and A. McCallum (2009). Rethinking LDA: Why priors matter. In *NIPS*.
- Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno (2009). Evaluation methods for topic models. In *ICML*.
- Williamson, S., C. Wang, K. A. Heller, and D. M. Blei (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*.
- Wolpert, R. L., M. A. Clyde, and C. Tu (2011). Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. *Ann. Statist.* 39(4), 1916–1962.
- Zhou, M. (2014). Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling. In *NIPS*, pp. 3455–3463.
- Zhou, M. and L. Carin (2015). Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(2), 307–320.
- Zhou, M., Y. Cong, and B. Chen (2015). The Poisson gamma belief network. In *NIPS*.
- Zhou, M., L. Hannah, D. Dunson, and L. Carin (2012). Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pp. 1462–1471.
- Zhou, M., O. H. M. Padilla, and J. G. Scott (2016). Priors for random count matrices derived from a family of negative binomial processes. *to appear in J. Amer. Statist. Assoc.*

# Nonparametric Bayesian Negative Binomial Factor

## Analysis: Supplementary Material

### A Proofs

*Proof of Theorem 1.* With  $\mathbf{t} := (t_0, \dots, t_K) \in \mathbb{R}^{K+1}$ , the characteristic function of  $\mathbf{x}$  can be expressed as

$$\mathbb{E} \left[ e^{i\mathbf{t}^T \mathbf{x}} \right] = \prod_{k=1}^K \mathbb{E} \left[ e^{i(t_0+t_k)x_k} \right] = \prod_{k=1}^K \left( \frac{1-p}{1-pe^{i(t_0+t_k)}} \right)^{r_k}. \quad (\text{A.1})$$

We augment  $\mathbf{y}$  as

$$(y_1, \dots, y_K) \sim \text{Mult}(\mathbf{y}, \boldsymbol{\theta}), \quad \boldsymbol{\theta} \sim \text{Dir}(r_1, \dots, r_K), \quad y \sim \text{Pois}(\lambda), \quad \lambda \sim \text{Gamma}(r, p/(1-p)) \quad (\text{A.2})$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^T$ . Conditioning on  $\boldsymbol{\theta}$  and  $y$ , we have

$$\mathbb{E} \left[ e^{i\mathbf{t}^T \mathbf{y}} \mid \boldsymbol{\theta}, y \right] = \left( \sum_{k=1}^K \theta_k e^{i(t_0+t_k)} \right)^y;$$

conditioning on  $\boldsymbol{\theta}$  and  $\lambda$ , we have

$$\begin{aligned} \mathbb{E} \left[ e^{i\mathbf{t}^T \mathbf{y}} \mid \boldsymbol{\theta}, \lambda \right] &= \mathbb{E}_y \left[ \left( \sum_{k=1}^K \theta_k e^{i(t_0+t_k)} \right)^y \right] \\ &= \exp \left[ \lambda \left( \sum_{k=1}^K \theta_k e^{i(t_0+t_k)} - 1 \right) \right] \\ &= \prod_{k=1}^K e^{\lambda_k (e^{i(t_0+t_k)} - 1)}, \end{aligned} \quad (\text{A.3})$$

where  $\lambda_k = \lambda \theta_k$  are independent gamma random variables, as the independent product of the gamma random variable  $\lambda$  and the Dirichlet random vector  $\boldsymbol{\theta}$ , with the gamma shape parameter and Dirichlet concentration parameter both equal to  $r$ , leads to independent gamma random variables. Further marginalizing out  $\lambda_k$ , we have

$$\mathbb{E} \left[ e^{i\mathbf{t}^T \mathbf{y}} \right] = \left[ 1 - \frac{p}{1-p} \left( e^{i(t_0+t_k)} - 1 \right) \right]^{-r_k} = \prod_{k=1}^K \left( \frac{1-p}{1-pe^{i(t_0+t_k)}} \right)^{r_k}. \quad (\text{A.4})$$

Thus  $\mathbf{x}$  and  $\mathbf{y}$  are equal in distribution as their characteristic functions are the same.  $\square$

*Poof of Proposition 2.* Multiplying the likelihood in (13) with the PMF of the NB distribution in (14), we have

$$P(\mathbf{x}_j, \mathbf{z}_j, n_j \mid \Phi, \theta_j, p_j) = \frac{1}{n_j!} \prod_{v=1}^V \prod_{k=1}^K \frac{\Gamma(n_{vjk} + \phi_{vk}\theta_{kj})}{\Gamma(\phi_{vk}\theta_{kj})} p_j^{n_{vjk}} (1-p_j)^{\phi_{vk}\theta_{kj}}, \quad (\text{A.5})$$

which, multiplied by the combinatorial coefficient  $n_j! / (\prod_{v=1}^V \prod_{k=1}^K n_{vjk}!)$ , becomes the same as (15).  $\square$

*Proof of Theorem 3.* For the first hierarchical model, we have

$$\begin{aligned} P(\mathbf{n}, \ell \mid n, \mathbf{r}) &= \left\{ \prod_{k=1}^K \text{CRT}(\ell_k; n_k, r_k) \right\} \text{DirMult}(\mathbf{n}; n, \mathbf{r}) \\ &= \frac{n!}{\prod_{k=1}^K n_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^K r_k^{\ell_k} |s(n_k, \ell_k)| \\ &= \frac{n!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^K \left\{ r_k^{\ell_k} \sum_{(n_{k1}, \dots, n_{k\ell_k}) \in \mathcal{D}_{n_k, \ell_k}} \prod_{t=1}^{\ell_k} \frac{1}{n_{kt}} \right\}. \end{aligned} \quad (\text{A.6})$$

Summing over all  $\mathbf{n}$  in the set  $\mathcal{M}_{n,K} = \{(n_1, \dots, n_K) : n_k \geq 0 \text{ and } \sum_{k=1}^K n_k = n\}$ , we have

$$\begin{aligned}
P(\boldsymbol{\ell} | n, \mathbf{r}) &= \sum_{(n_1, \dots, n_K) \in \mathcal{M}_{n,K}} \frac{n!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^K \left\{ r_k^{\ell_k} \sum_{(n_{k1}, \dots, n_{k\ell_k}) \in \mathcal{D}_{n_k, \ell_k}} \prod_{t=1}^{\ell_k} \frac{1}{n_{kt}} \right\} \\
&= \left\{ \frac{n!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^K r_k^{\ell_k} \right\} \left\{ \sum_{(n_1, \dots, n_K) \in \mathcal{M}_{n,K}} \sum_{(n_{k1}, \dots, n_{k\ell_k}) \in \mathcal{D}_{n_k, \ell_k}} \prod_{t=1}^{\ell_k} \frac{1}{n_{kt}} \right\} \\
&= \left\{ \frac{n!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^K r_k^{\ell_k} \right\} \left\{ \sum_{(n_1, \dots, n_{\ell.}) \in \mathcal{D}_{n, \ell.}} \prod_{t=1}^{\ell.} \frac{1}{n_t} \right\} \\
&= \frac{\ell. !}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r.)}{\Gamma(n+r.)} |s(n, \ell.)| \prod_{k=1}^K r_k^{\ell_k} \\
&= \text{Mult}(\boldsymbol{\ell}; \ell., r_1/r., \dots, r_K/r.) \text{CRT}(\boldsymbol{\ell}; n, r.) \tag{A.7}
\end{aligned}$$

□

## B Posterior analysis for the hGNBP

Denote  $L \sim \text{CRTP}(X, G)$  as a CRT process such that  $L(A) = \sum_{\omega \in A} L(\omega)$ ,  $L(\omega) \sim \text{CRT}[X(\omega), G(\omega)]$  for each  $A \subset \Omega$ , and  $X \sim \text{SumLogP}(L, p)$  as a sum-logarithmic process such that  $X(A) \sim \text{SumLog}[L(A), p]$  for each  $A \subset \Omega$ . As in Zhou and Carin (2015), generalizing the Poisson-logarithmic bivariate distribution in Section 2.1, one may show that  $X$  and  $L$  given  $G$  and  $p$  in

$$L | X \sim \text{CRTP}(X, G), \quad X \sim \text{NBP}(G, p)$$

is equivalent in distribution to those in

$$X | L \sim \text{SumLogP}(L, p), \quad L \sim \text{PP}[-G \ln(1-p)],$$

where  $L \sim \text{PP}[-G \ln(1-p)]$  is a Poisson process such that  $L(A) \sim \text{Pois}[-G(A) \ln(1-p)]$  for each  $A \subset \Omega$ . Generalizing the analysis for the GNBP in Zhou and Carin (2015) and Zhou

et al. (2016), with

$$\tilde{p}_j := \frac{-\ln(1-p_j)}{c_j - \ln(1-p_j)}, \quad \tilde{\tilde{p}} := \frac{-\sum_j \ln(1-\tilde{p}_j)}{c_0 - \sum_j \ln(1-\tilde{p}_j)},$$

we can express the conditional posteriors of  $G$  and  $\Theta_j$  as

$$\begin{aligned} (L_j | X_j, \Theta_j) &\sim \text{CRTP}(X_j, \Theta_j), \quad (\tilde{L}_j | L_j, G) \sim \text{CRTP}(L_j, G), \\ (G | \{\tilde{L}_j, p_j\}_j, G_0) &\sim \Gamma\text{P}\left(G_0 + \sum_j \tilde{L}_j, [c_0 - \sum_j \ln(1-\tilde{p}_j)]^{-1}\right), \\ (\Theta_j | G, \tilde{L}_j, p_j, c_j) &\sim \Gamma\text{P}(G + L_j, [c_j - \ln(1-p_j)]^{-1}). \end{aligned} \quad (\text{B.1})$$

If we let  $\gamma_0 \sim \text{Gamma}(a_0, 1/b_0)$ , the conditional posterior of  $\gamma_0$  can be expressed as

$$\begin{aligned} (\tilde{\tilde{L}} | \{\tilde{L}_j\}_j, G_0) &\sim \text{CRTP}(\sum_j \tilde{L}_j, G_0), \\ (\gamma_0 | \tilde{\tilde{L}}, \{p_j, c_j\}_j, c_0) &\sim \text{Gamma}\left(a_0 + \tilde{\tilde{L}}(\Omega), [b_0 - \ln(1-\tilde{\tilde{p}})]^{-1}\right). \end{aligned} \quad (\text{B.2})$$

If the base measure  $G_0$  is finite and continuous, we have

$$\tilde{\tilde{L}}(\Omega) = \sum_{k=1}^{\infty} \delta\left(\sum_j \tilde{L}_j(\phi_k) > 0\right) = \sum_{k=1}^{\infty} \delta\left(\sum_j X_j(\phi_k) > 0\right),$$

which is the number of active atoms that are associated with nonzero counts, otherwise we have  $(\tilde{\tilde{L}}(\omega_k) | \{\tilde{L}_j\}_j, G_0) \sim \text{CRT}[\sum_j \tilde{L}_j(\omega_k), G_0(\omega_k)]$  for all atoms  $\omega_k \in \Omega$ . In this paper, we let  $K^+ = \tilde{\tilde{L}}(\Omega)$  denote the number of active atoms.

# C Gibbs sampling for the hGNBP Dirichlet-multinomial mixed-membership model

## C.1 Blocked Gibbs sampling

As it is impossible to instantiate all the countably infinite atoms of a gamma process draw, expressed as  $G = \sum_{k=1}^{\infty} r_k \delta_{\phi_k}$ , for the convenience of implementation, it is common to consider truncating the total number of atoms to be  $K$  by choosing a discrete base measure as  $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K} \delta_{\phi_k}$ , under which we have  $r_k \sim \text{Gamma}(\gamma_0/K, 1/c_0)$  for  $k \in \{1, \dots, K\}$  (Zhou and Carin, 2015). The finite truncation strategy is also commonly used for the Dirichlet process based mixture models (Neal, 2000; Ishwaran and James, 2001; Fox et al., 2011) and the beta process (Hjort, 1990) based factor models (Zhou et al., 2012). Although fixing the truncation level  $K$  often works well in practice, it may lead to a considerable waste of computation if  $K$  is set to be too large. One may also consider instantiating only the atoms whose weights are larger than a predefined threshold and using reversible jump MCMC to

---

**Algorithm 1** Gibbs sampling algorithms for the hierarchical gamma-negative binomial process negative binomial factor analysis (Dirichlet-multinomial mixed-membership model).

---

```

1: for  $iter = 1 : MaxIter$  do Gibbs sampling
2:   switch Gibbs sampler do
3:     case regular blocked Gibbs sampler
4:       sample  $\{z_{ji}\}_{j,i}$  and then calculate  $\{n_{vjk}\}_{v,j,k}$ ; sample a latent count  $\ell_{vjk}$  for each  $n_{vjk}$ ;
5:     case collapsed Gibbs sampler
6:       sample  $\{z_{ji}, b_{ji}\}_{j,i}$  and then calculate  $\{n_{vjk}, \ell_{vjk}\}_{v,j,k}$ ;
7:     case compound Poisson based blocked Gibbs sampler
8:       sample a latent count  $\ell_{vj}$  for each  $n_{vj}$ ; sample  $\{\ell_{vjk}\}_k$  for each  $\ell_{vj}$ ;
9:   end switch
10:  sample  $\{p_j\}_j$ ; sample  $\{c_j\}_j$ ; sample  $\gamma_0$ ; sample  $c_0$ ; relabel the active factors from 1 to  $K^+$ .
11:  if collapsed Gibbs sampler then
12:    sample  $\{r_k\}_{1,K^+}$ ; sample  $r_\star$ ; sample  $\{\theta_{.j}\}_j$ ;
13:  else
14:    for  $k = 1 : K^+ + K_\star$  do
15:      sample  $\phi_k^{(t)}$ ; sample  $r_k$ ; sample  $\{\theta_{kj}\}_j$ ;
16:    end for
17:  end if
18: end for

```

---

infer the number of active atoms (Wolpert et al., 2011).

For nonparametric Bayesian mixture models based on the Dirichlet process (Ferguson, 1973; Escobar and West, 1995) or other more general normalized random measures with independent increments (NRMIs) (Regazzini et al., 2003; Lijoi et al., 2007; Lijoi and Prünster, 2010; Favaro and Teh, 2013), one may consider slice sampling to adaptively truncated the number of atoms used in each Markov chain Monte Carlo (MCMC) iteration (Walker, 2007; Papaspiliopoulos and Roberts, 2008; Griffin and Walker, 2011). Unlike NRMIs whose atoms' weights have to sum to one and hence are negatively correlated, the weights of the atoms of completely random measures are independent from each other. Exploiting this property, for our models built on completely random measures, we construct a sampling procedure that adaptively truncates the total number of atoms in each iteration.

Note that the conditional posterior of  $G$  shown in (B.1) can be written as the summation of two independent gamma processes:  $\mathcal{D} \sim \Gamma\text{P}(\sum_j \tilde{L}_j, [c_0 - \sum_j \ln(1 - \tilde{p}_j)]^{-1})$  and  $G_\star \sim \Gamma\text{P}(G_0, [c_0 - \sum_j \ln(1 - \tilde{p}_j)]^{-1})$ . To approximately represent a draw from

$$G_\star \sim \Gamma\text{P}\left(G_0, [c_0 - \sum_j \ln(1 - \tilde{p}_j)]^{-1}\right)$$

that consists of countably infinite atoms, at the end of each MCMC iteration, we relabel the indices of the atoms with nonzero counts from 1 to  $K^+ := \sum_{k=1}^{\infty} \delta(\sum_j X_j(\phi_k) > 0)$ ; instantiate  $K_\star$  new atoms as

$$\tilde{G}_\star = \sum_{k=K^++1}^{K^++K_\star} r_k \delta_{\phi_k}, \quad r_k \sim \text{Gamma}\left(\frac{\gamma_0}{K_\star}, \frac{1}{c_0 - \sum_j \ln(1 - \tilde{p}_j)}\right), \quad \phi_k \sim \text{Dir}(\eta, \dots, \eta);$$

and set  $K := K^+ + K_\star$  as the total number of atoms to be used for the next iteration.

We outline in Algorithm 1 the blocked Gibbs sampler and the other two Gibbs samplers for the hGNBP, and present the update equations below. Each blocked Gibbs sampling iteration for the hGNBP DMMM model in (23) proceeds as follows.

**Sample**  $z_{ji}$ . Using the likelihood in (13), we have

$$P(z_{ji} = k | x_{ji}, \mathbf{z}_j^{-i}, \Phi, \theta_j) \propto n_{x_{ji}jk}^{-ji} + \phi_{x_{ji}k} \theta_{kj}, \quad k \in \{1, \dots, K\}. \quad (\text{C.1})$$

**Sample**  $\ell_{vjk}$ . Since  $n_{vjk} \sim \text{NB}(\phi_{vk} \theta_{kj}, p_j)$  in the prior, as shown in Proposition 2, we can draw a corresponding latent count  $\ell_{vjk}$  for each  $n_{vjk}$  as

$$(\ell_{vjk} | -) \sim \text{CRT}(n_{vjk}, \phi_{vk} \theta_{kj}), \quad (\text{C.2})$$

where  $\ell_{vjk} = 0$  almost surely if  $n_{vjk} = 0$ .

**instance**  $p_j$ . We sample  $p_j$  as

$$(p_j | -) \sim \text{Beta}(a_0 + n_j, b_0 + \theta_{.j}). \quad (\text{C.3})$$

**Sample**  $c_j$ . We sample  $c_j$  as

$$(c_j | -) \sim \text{Gamma}[e_0 + G(\Omega), 1/(f_0 + \theta_{.j})]. \quad (\text{C.4})$$

**Sample**  $r_k$ . We first sample latent counts and then sample  $\gamma_0$  and  $c_0$  as

$$\begin{aligned} (\tilde{\ell}_{jk} | -) &\sim \text{CRT}(\ell_{.jk}, r_k), \quad (\gamma_0 | -) \sim \text{Gamma}\left(a_0 + K^+, \frac{1}{b_0 - \ln(1 - \tilde{p})}\right), \\ (c_0 | -) &\sim \text{Gamma}\left(e_0 + \gamma_0, \frac{1}{f_0 + G(\Omega)}\right), \end{aligned} \quad (\text{C.5})$$

where  $\tilde{\ell}_{.k} := \sum_j \tilde{\ell}_{jk}$ ,  $K^+ := \sum_k \delta(n_{..k} > 0) = \sum_k \delta(\tilde{\ell}_{.k} > 0)$ . For all the points of discontinuity, *i.e.*, the factors in the set  $\{\phi_k\}_{k:n_{..k}>0}$ , we relabel their indices from 1 to  $K^+$  and then sample  $r_k$  as

$$(r_k | -) \sim \text{Gamma}\left(\tilde{\ell}_{.k}, \frac{1}{c_0 - \sum_j \ln(1 - \tilde{p}_j)}\right), \quad (\text{C.6})$$

and for the absolutely continuous space  $\{\phi_k\}_{k:n_{..k}=0}$ , we instantiate  $K_\star$  unused atoms, whose



weights are sampled as

$$(r_k | -) \sim \text{Gamma}\left(\frac{\gamma_0}{K_\star}, \frac{1}{c_0 - \sum_j \ln(1 - \tilde{p}_j)}\right). \quad (\text{C.7})$$

We let  $K := K^+ + K_\star$  and  $G(\Omega) := \sum_{k=1}^K r_k$ .

**instance**  $\phi_k$ . Denote  $\ell_{v \cdot k} = \sum_{j=1}^J \ell_{vjk}$ . Since  $\ell_{vjk} \sim \text{Pois}[-\phi_{vk} \theta_{kj} \ln(1 - p_j)]$  in the prior, we can sample  $\phi_k$  as

$$(\phi_k | -) \sim \text{Dir}(\eta + \ell_{1 \cdot k}, \dots, \eta + \ell_{V \cdot k}). \quad (\text{C.8})$$

**instance**  $\theta_{kj}$ . Denote  $\ell_{ \cdot jk} = \sum_{v=1}^V \ell_{vjk}$ . We can sample  $\theta_{kj}$  as

$$(\theta_{kj} | -) \sim \text{Gamma}[r_k + \ell_{ \cdot jk}, 1/(c_j - \ln(1 - p_j))]. \quad (\text{C.9})$$

## C.2 Collapsed Gibbs sampling

One common strategy to improve convergence and mixing for multinomial mixed-membership models is to collapse the factors  $\{\phi_k\}_k$  and factor scores  $\{\theta_j\}_j$  in the sampler (Griffiths and Steyvers, 2004; Newman et al., 2009). To apply this strategy to the DMMM model, we first need to transform the likelihood in (18) to make it amenable to marginalization. Using an analogy similar to that for the Chinese restaurant franchise of Teh et al. (2006), if we consider  $z_{ji}$  as the index of the “dish” that the  $i$ th “customer” in the  $j$ th “restaurant” takes, then, to make the likelihood in (18) become fully factorized, we may introduce  $b_{ji}$  as the index of the table at which this customer is seated. The following proposition reveals how the Chinese restaurant process can be related to the Dirichlet-multinomial and NB distributions, and shows how to introduce auxiliary variables to make the likelihood of  $\mathbf{z} \sim \text{DirCat}(n, r_1, \dots, r_K)$ , as shown in (2), become fully factorized.

**Proposition 4.** *Given the instance length  $n$  (number of customers) and  $\mathbf{r} = (r_1, \dots, r_K)$ , the joint distribution of the “table” indices  $\mathbf{b} = (b_1, \dots, b_n)$  and “dish” indices  $\mathbf{z} = (z_1, \dots, z_n)$*

in

$$\{b_i\}_{i:z_i=k} \sim \text{CRP}(n_k, r_k), \quad \mathbf{z} \sim \text{DirCat}(n, r_1, \dots, r_K),$$

is the same as that in

$$z_i = s_{b_i}, \quad s_t \sim \text{Cat}(r_1/r, \dots, r_K/r), \quad \mathbf{b} \sim \text{CRP}(n, r),$$

with PMF

$$P(\mathbf{b}, \mathbf{z} \mid n, \mathbf{r}) = \frac{\Gamma(r.)}{\Gamma(n + r.)} \prod_{k=1}^K \left\{ r_k^{\ell_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\},$$

where  $\ell_k$  is the number of unique indices in  $\{b_i\}_{i:z_i=k}$ ,  $\ell. = \sum_{k=1}^K \ell_k$  is the total number of nonempty tables,  $t = 1, \dots, \ell.$ , and  $n_{kt} = \sum_{i=1}^n \delta(b_i = t, z_i = k)$  is the number of customers that sit at table  $t$  and take dish  $k$ .

If we further randomize the instance length as

$$n \sim \text{NB}(r., p),$$

then we have the PMF for the joint distribution of  $\mathbf{b}$ ,  $\mathbf{z}$ , and  $n$  given  $\mathbf{r}$  and  $p$  in a fully factorized form as

$$P(\mathbf{b}, \mathbf{z}, n \mid \mathbf{r}, p) = \frac{1}{n!} \prod_{k=1}^K \left\{ r_k^{\ell_k} (1-p)^{r_k} p^{n_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\},$$

which, with appropriate combinatorial analysis, can be mapped to the PMF of the joint distribution of  $\mathbf{n} = (n_1, \dots, n_K)$ ,  $\boldsymbol{\ell} = (\ell_1, \dots, \ell_K)$ , and  $n$  given  $\mathbf{r}$  and  $p$  in

$$n = \sum_{k=1}^K n_k, \quad n_k = \sum_{t=1}^{\ell_k} n_{kt}, \quad n_{kt} \sim \text{Log}(p), \quad \ell_k \sim \text{Pois}[-r_k \ln(1-p)]. \quad (\text{C.10})$$

*Proof.* For the first hierarchical model, we have

$$\begin{aligned}
P(\mathbf{b}, \mathbf{z} \mid n, \mathbf{r}, p) &= \left\{ \prod_{k=1}^K \text{CRP}(\{b_i\}_{i:z_i=k}; n_k, r_k) \right\} \text{DirCat}(\mathbf{z}; n, \mathbf{r}) \\
&= \frac{\Gamma(r.)}{\Gamma(n + r.)} \prod_{k=1}^K \text{CRP}(\{b_i\}_{i:z_i=k}; n_k, r_k) \frac{\Gamma(n_k + r_k)}{\Gamma(r_k)} \\
&= \frac{\Gamma(r.)}{\Gamma(n + r.)} \prod_{k=1}^K \left\{ r_k^{\ell_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\}, \tag{C.11}
\end{aligned}$$

where  $\ell_k$  is the number of unique indices in  $\{b_i\}_{i:z_i=k}$  and  $n_{kt} = \sum_{i=1}^n \delta(b_i = t, z_i = k)$ .

For the second hierarchical model, we have

$$\begin{aligned}
P(\mathbf{b}, \mathbf{z} \mid n, \mathbf{r}, p) &= P(\mathbf{z} \mid \mathbf{b}, \mathbf{r}) \text{CRP}(\mathbf{b}; n, r.) \\
&= P(\mathbf{z} \mid \mathbf{b}, \mathbf{r}) r.^\ell \frac{\Gamma(r.)}{\Gamma(n + r.)} \prod_{t=1}^{\ell} \Gamma\left(\sum_{i=1}^n \delta(b_i = t)\right) \\
&= \frac{\Gamma(r.)}{\Gamma(n + r.)} \left\{ \prod_{k=1}^K r_k^{\sum_{t=1}^{\ell_k} \delta(s_t=k)} \right\} \left\{ \prod_{t=1}^{\ell} \Gamma\left(\sum_{i=1}^n \delta(b_i = t)\right) \right\} \\
&= \frac{\Gamma(r.)}{\Gamma(n + r.)} \prod_{k=1}^K \left\{ r_k^{\ell_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\}, \tag{C.12}
\end{aligned}$$

where  $\ell_k = \sum_{t=1}^{\ell} \delta(s_t = k)$  is the number of unique indices in  $\{b_i\}_{i:z_i=k}$  and  $n_{kt} = \delta(s_t = k) \sum_{i=1}^n \delta(b_i = t) = \sum_{i=1}^n \delta(b_i = t, z_i = k)$ .

Simply applying the chain rule, we have

$$P(\mathbf{b}, \mathbf{z}, n \mid \mathbf{r}, p) = P(\mathbf{b}, \mathbf{z} \mid n, \mathbf{r}) \text{NB}(n; r., p) = \frac{1}{n!} \prod_{k=1}^K \left\{ r_k^{\ell_k} (1-p)^{r_k} p^{n_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\} \tag{C.13}$$

The mapping from  $\{\mathbf{b}, \mathbf{z}, n\}$  to  $\{\boldsymbol{\ell}, \mathbf{n}, n\}$  is many to one, with

$$\begin{aligned} P(\boldsymbol{\ell}, \mathbf{n}, n \mid \mathbf{r}, p) &= \prod_{k=1}^K \left\{ \sum_{(n_1, \dots, n_k) \in \mathcal{D}_{n_k, \ell_k}} \frac{1}{\ell_k! \prod_{t=1}^{\ell_k} n_{kt}!} r_k^{\ell_k} (1-p)^{r_k} p^{n_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\} \\ &= \prod_{k=1}^K \left\{ r_k^{\ell_k} (1-p)^{r_k} p^{n_k} \frac{|s(n_k, \ell_k)|}{n_k!} \right\}, \end{aligned} \quad (\text{C.14})$$

where  $\mathcal{D}_{n_k, \ell_k} := \{(n_{k1}, \dots, n_{k\ell_k}) : n_{kt} \geq 1 \text{ and } \sum_{t=1}^{\ell_k} n_{kt} = n_k\}$ . For (C.10), using the PMF of the Poisson-logarithmic bivariate distribution shown in (1), we have

$$P(\boldsymbol{\ell}, \mathbf{n}, n \mid \mathbf{r}, p) = \prod_{k=1}^K \left\{ r_k^{\ell_k} (1-p)^{r_k} p^{n_k} \frac{|s(n_k, \ell_k)|}{n_k!} \right\}. \quad (\text{C.15})$$

□

Using (18) and Proposition 4, introducing the auxiliary variables

$$\{b_{ji}\}_{i: x_{ji}=v, z_{ji}=k} \sim \text{CRP}(n_{vjk}, \phi_{vk} \theta_{kj}), \quad (\text{C.16})$$

we have the joint likelihood for the DMMM model as

$$P(\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j) = \frac{1}{n_j!} \prod_v \prod_k \left\{ (\phi_{vk} \theta_{kj})^{\ell_{vjk}} p_j^{n_{vjk}} (1-p_j)^{\phi_{vk} \theta_{kj}} \prod_{t=1}^{\ell_{vjk}} \Gamma(n_{vjk t}) \right\}, \quad (\text{C.17})$$

where  $\ell_{vjk}$  is the number of unique indices in  $\{b_{ji}\}_{i: x_{ji}=v, z_{ji}=k}$  and  $n_{vjk t} = \sum_{i=1}^{n_j} \delta(x_{ji} = v, z_{ji} = k, b_{ji} = t)$ . As in Proposition 4, instead of first assigning the attribute tokens to factors using the Dirichlet-multinomial distributions and then assigning the attribute tokens with the same factor indices to tables, we may first assign the attribute tokens to tables and

then assign the tables to factors. Thus we have the following model

$$\begin{aligned} x_{ji} &= w_{jz_{ji}}, \quad z_{ji} = s_{jb_{ji}}, \quad w_{js_{jt}} \sim \text{Cat}(\boldsymbol{\phi}_{s_{jt}}), \quad s_{jt} \sim \text{Cat}(\boldsymbol{\theta}_j/\theta_{.j}), \\ \mathbf{b}_j &\sim \text{CRP}(n_j, \theta_{.j}), \quad n_j \sim \text{NB}(\theta_{.j}, p_j), \end{aligned} \quad (\text{C.18})$$

whose likelihood  $P(\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j)$  is the same as the likelihood, as shown in (C.17), of the DMMM model constituted of (12) and (14) and augmented with (C.16).

We outline the collapsed Gibbs sampler for the hGNBP-NBFA in Algorithm 1 and provide the derivation and update equations below. This collapsed sampling strategy marginalizes out both the factors  $\{\boldsymbol{\phi}_k\}$  and factor scores  $\{\boldsymbol{\theta}_j\}_j$ , but at the expense of introducing an auxiliary variable  $b_{ji}$  for each attribute token  $x_{ji}$ .

Marginalizing out  $\boldsymbol{\Phi}$  and  $\boldsymbol{\Theta}$  from  $\prod_j P(\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j)$ , we have

$$\begin{aligned} P(\{\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j\}_j \mid G, \mathbf{p}) &= e^{r_\star \sum_j \ln(1-p_j)} \left\{ \prod_j p_j^{n_j} \frac{\prod_v \prod_k \left( \prod_{t=1}^{\ell_{vjk}} \Gamma(n_{vjkt}) \right)}{n_j!} \right\} \\ &\times \left\{ \prod_{k: \ell_{..k} > 0} \frac{\Gamma(V\eta)}{\Gamma(\ell_{..k} + V\eta)} \prod_{v=1}^V \frac{\Gamma(\ell_{v..k} + \eta)}{\Gamma(\eta)} \right\} \\ &\times \left\{ \prod_j \prod_{k: \ell_{..k} > 0} \frac{\Gamma(r_k + \ell_{.jk})}{\Gamma(r_k)} \frac{c_j^{r_k}}{[c_j - \ln(1-p_j)]^{r_k + \ell_{.jk}}} \right\}, \quad (\text{C.19}) \end{aligned}$$

where  $\ell_{..k} := \sum_j \ell_{.jk}$  and  $r_\star := \sum_{k: \ell_{..k} = 0} r_k$ .

**Sample  $z_{ji}$  and  $b_{ji}$ .** Using the likelihood in (C.19), with  $(K^+)^{-ji}$  representing the number

of active atoms without considering  $z_{ji}$ , we have

$$P(z_{ji} = k, b_{ji} = t \mid x_{ji}, \mathbf{z}^{-ji}, \mathbf{b}^{-ji}, G) \propto \begin{cases} n_{x_{ji}jkt}^{-ji}, & \text{if } k \leq (K^+)^{-ji}, t \leq \ell_{x_{ji}jk}^{-ji}; \\ \frac{\ell_{x_{ji}\cdot k}^{-ji} + \eta}{\ell_{\cdot\cdot k}^{-ji} + V\eta} \frac{r_k + \ell_{\cdot jk}^{-ji}}{c_j - \ln(1 - p_j)}, & \text{if } k \leq (K^+)^{-ji}, t = \ell_{x_{ji}jk}^{-ji} + 1; \\ \frac{1}{V} \frac{r_\star}{c_j - \ln(1 - p_j)}, & \text{if } k = (K^+)^{-ji} + 1, t = 1; \end{cases} \quad (\text{C.20})$$

and if  $k = (K^+)^{-ji} + 1$  happens, similar to the direct assignment sampler for the hierarchical Dirichlet process (Teh et al., 2006), we draw  $\beta \sim \text{Beta}(1, \gamma_0)$  and then let  $r_k = \beta r_\star$  and  $r_\star = (1 - \beta)r_\star$ . This is based on the stick-breaking representation of the Dirichlet process,  $\tilde{G}_\star \sim \text{DP}(\gamma_0, G_0/\gamma_0)$ , whose product with an independent random variable  $r_\star \sim \left(\gamma_0, [c_0 - \sum_j \ln(1 - \tilde{p}_j)]^{-1}\right)$  recovers the gamma process  $G_\star \sim \Gamma\text{P}\left(G_0, [c_0 - \sum_j \ln(1 - \tilde{p}_j)]^{-1}\right)$ . Note that instead of drawing both  $z_{ji}$  and  $b_{ji}$  at the same time, one may first draw  $z_{ji}$  and then draw  $b_{ji}$  given  $z_{ji}$ , or first draw  $b_{ji}$  and then draw  $z_{ji}$  given  $b_{ji}$ .

The other model parameters can all be sampled in the way similar to how they are sampled in Section C.1. Below we highlight the differences. First, we do not need to sample  $\{\phi_k\}$ . Instead of sampling  $\{\theta_{kj}\}_k$ , we only need to sample  $\theta_{\cdot j}$  as

$$(\theta_{\cdot j} \mid -) \sim \text{Gamma}[G(\Omega) + \sum_k \ell_{\cdot jk}, 1/(c_j - \ln(1 - p_j))]. \quad (\text{C.21})$$

For the absolutely continuous space, we have

$$(r_\star \mid -) \sim \text{Gamma}\left(\gamma_0, \frac{1}{c_0 - \sum_j \ln(1 - \tilde{p}_j)}\right). \quad (\text{C.22})$$

We have  $K^+ = \sum_k \delta(\ell_{\cdot\cdot k} > 0)$  and  $G(\Omega) := r_\star + \sum_{k: \ell_{\cdot\cdot k} > 0} r_k$ .

## D Gamma-negative binomial process PFA and DCMLDA

We consider the GNB (Zhou and Carin, 2015) as

$$X_j \sim \text{NB}(G, p_j), \quad G \sim \text{GP}(G_0, 1/c_0). \quad (\text{D.1})$$

The GNB multinomial mixed-membership model of Zhou and Carin (2015) can be expressed as

$$\begin{aligned} x_{ji} &\sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}), \quad z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j / \theta_{\cdot j}), \\ \theta_{kj} &\sim \text{Gamma}[r_k, p_j / (1 - p_j)], \\ n_j &\sim \text{Pois}(\theta_{\cdot j}), \quad p_j \sim \text{Beta}(a_0, b_0), \\ G &\sim \text{GP}(G_0, 1/c_0), \end{aligned} \quad (\text{D.2})$$

which, as far as the conditional posteriors of  $\{\boldsymbol{\phi}_k\}_k$  and  $\{\boldsymbol{\theta}_j\}_j$  are concerned, can be equivalently represented as the GNB-PFA

$$\begin{aligned} n_{vj} &= \sum_{k=1}^{\infty} n_{vjk}, \quad n_{vjk} \sim \text{Pois}(\phi_{vk} \theta_{kj}), \\ \theta_{kj} &\sim \text{Gamma}[r_k, p_j / (1 - p_j)], \\ p_j &\sim \text{Beta}(a_0, b_0), \quad G \sim \text{GP}(G_0, 1/c_0). \end{aligned} \quad (\text{D.3})$$

Similar to how adaptive truncation is used in blocked Gibbs sampling for the hGNB-NBFA, one may readily extend the blocked Gibbs sampler for the GNB multinomial mixed-membership model developed in Zhou and Carin (2015), which has a fixed finite truncation, to a one with adaptive truncation. We omit these details for brevity. We describe a collapsed Gibbs sampler for the GNB-PFA in Appendix D.1.

As discussed before, the GNB can also be applied to DCMLDA to support countably infinite factors. We express the GNB-DCMLDA as

$$\begin{aligned}
x_{ji} &\sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}), \quad z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j), \\
\boldsymbol{\phi}_k^{[j]} &\sim \text{Dir}(\boldsymbol{\phi}_k r_k), \quad \boldsymbol{\theta}_j \sim \text{Dir}(\mathbf{r}), \\
n_j &\sim \text{NB}(r_., p_j), \quad p_j \sim \text{Beta}(a_0, b_0), \\
G &\sim \Gamma\text{P}(G_0, 1/c_0),
\end{aligned} \tag{D.4}$$

which, as far as the conditional posteriors of  $\{\boldsymbol{\phi}_k\}_k$  and  $\{r_k\}_k$  are concerned, can be equivalently represented as

$$\begin{aligned}
n_{vj} &= \sum_{k=1}^{\infty} n_{vjk}, \quad n_{vjk} \sim \text{NB}(\phi_{vk} r_k, p_j). \\
p_j &\sim \text{Beta}(a_0, b_0), \quad G \sim \Gamma\text{P}(G_0, 1/c_0).
\end{aligned} \tag{D.5}$$

The restriction is evident from (D.5) as all the instances are enforced to have the same factor scores as  $\mathbf{r}_k$  under the shared factors  $\{\boldsymbol{\phi}_k\}_k$ . Blocked Gibbs sampling with and without sampling  $z_{ji}$  for the GNBP-DCMLDA can be similarly derived as those for the hGNBP DMMM model, omitted here for brevity. We describe in detail a collapsed Gibbs sampler for the GNBP-DCMLDA in Appendix D.2.

## D.1 Collapsed Gibbs sampling for GNB-PFA

For the GNB in (D.1), the conditional likelihood  $p(\{X_j\}_{1,J} | G)$  is shown in Appendix B.1 of Zhou et al. (2016). As there is a one-to-many mapping from  $\{X_j\}_{1,J}$  to  $\mathbf{z} = \{z_{11}, \dots, z_{Jm_J}\}$ , similar to the analysis in Zhou (2014), we have the joint likelihood of  $\mathbf{z}$  and the instance lengths  $\mathbf{m} = (m_1, \dots, m_J)$  as

$$p(\mathbf{z}, \mathbf{m} | G, \mathbf{p}) = \frac{p(\{X_j\}_{1,J} | G)}{\prod_{j=1}^J \frac{n_j!}{\prod_{k=1}^{\infty} n_{jk}!}} = \prod_{j=1}^J \frac{p_j^{n_j}}{n_j!} \prod_{k=1}^{\infty} \frac{\Gamma(n_{jk} + r_k)}{\Gamma(r_k)} (1 - p_j)^{r_k}. \tag{D.6}$$

Assuming the  $K^+$  factors that are associated with nonzero counts are relabeled in an arbitrary order from 1 to  $K^+$ , based on this conditional likelihood, we have a prediction rule



conditioning on  $G$  as

$$P(z_{ji} = k | \mathbf{z}^{-ji}, \mathbf{m}, G) \propto \begin{cases} n_{jk}^{-ji} + r_k, & \text{for } k = 1, \dots, (K^+)^{-ji}; \\ r_*, & \text{if } k = (K^+)^{-ji} + 1, \end{cases} \quad (\text{D.7})$$

where  $r_* = G(\Omega \setminus \{\phi_k\}_{1, K^+})$  is the total weight of all the factors assigned with zero count. This prediction rule becomes very similar to the direct assignment sampler of the hierarchical Dirichlet process (Teh et al., 2006) if one writes each  $r_k$  as the product of a total random mass  $\alpha$  and a probability  $\pi_k$ , with  $\alpha = \sum_{k=1}^{\infty} r_k$  and  $\sum_{k=1}^{\infty} \pi_k = 1$ . This is as expected since the gamma process can be represented as the independent product of a gamma process and a Dirichlet process, under the condition that the mass parameter of the gamma process is the same as the concentration parameter of the Dirichlet process, and the GNBP is closely related to the hierarchical Dirichlet process for mixed-membership modeling (Zhou and Carin, 2015).

Similar to the derivation of collapsed Gibbs sampling for the mixed-membership model based on the beta-negative binomial process, as shown in Zhou (2014), we can write the collapsed Gibbs sampling update equation for the factor indices as

$$P(z_{ji} = k | \mathbf{x}, \mathbf{z}^{-ji}, \mathbf{m}, G) \propto \begin{cases} \frac{\eta + n_{v_{ji} \cdot k}^{-ji}}{V\eta + n_{\cdot k}^{-ji}} \cdot (n_{jk}^{-ji} + r_k), & \text{for } k = 1, \dots, (K^+)^{-ji}; \\ \frac{1}{V} \cdot r_*, & \text{if } k = (K^+)^{-ji} + 1; \end{cases} \quad (\text{D.8})$$

and if  $k = (K^+)^{-ji} + 1$  happens, we draw  $\beta \sim \text{Beta}(1, \gamma_0)$  and then let  $r_k = \beta r_*$  and  $r_* = (1 - \beta)r_*$ . Gibbs sampling update equations for the other model parameters of the GNBP can be similarly derived as in Zhou and Carin (2015) and Zhou et al. (2016), omitted here for brevity.

## D.2 Collapsed Gibbs sampling for GNBP-DCMLDA

For collapsed Gibbs sampling of (D.4), introducing the auxiliary variables

$$\{b_{ji}\}_{i:x_{ji}=v, z_{ji}=k} \sim \text{CRP}(n_{vjk}, \phi_{vk} r_k), \quad (\text{D.9})$$

we have the joint likelihood of  $\mathbf{b}_j$ ,  $\mathbf{z}_j$ ,  $\mathbf{x}_j$  and  $n_j$  for DCMLDA as

$$P(\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j \mid \Phi, \theta_j, p_j) = \frac{1}{n_j!} \prod_v \prod_k \left\{ (\phi_{vk} r_k)^{\ell_{vjk}} p_j^{n_{vjk}} (1 - p_j)^{\phi_{vk} r_k} \prod_{t=1}^{\ell_{vjk}} \Gamma(n_{vjk t}) \right\}, \quad (\text{D.10})$$

where  $\ell_{vjk}$  is the number of unique indices in  $\{b_{ji}\}_{i:x_{ji}=v, z_{ji}=k}$  and  $n_{vjk t} = \sum_{i=1}^{n_j} \delta(x_{ji} = v, z_{ji} = k, b_{ji} = t)$ .

Marginalizing out  $\Phi$  from this likelihood, we have

$$\begin{aligned} P(\{\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j\}_j \mid G, \mathbf{p}) &= e^{r_\star \sum_j \ln(1-p_j)} \left\{ \prod_j p_j^{n_j} \frac{\prod_v \prod_k \left( \prod_{t=1}^{\ell_{vjk}} \Gamma(n_{vjk t}) \right)}{n_j!} \right\} \\ &\times \left\{ \prod_{k: \ell_{..k} > 0} r_k^{\ell_{..k}} e^{r_k \sum_j \ln(1-p_j)} \frac{\Gamma(V\eta)}{\Gamma(\ell_{..k} + V\eta)} \prod_{v=1}^V \frac{\Gamma(\ell_{v..} + \eta)}{\Gamma(\eta)} \right\}, \end{aligned} \quad (\text{D.11})$$

where  $r_\star := \sum_{k: \ell_{..k} > 0} r_k$ . With this likelihood, we have

$$P(z_{ji} = k, b_{ji} = t \mid x_{ji}, \mathbf{z}^{-ji}, \mathbf{b}^{-ji}, G) \propto \begin{cases} n_{x_{ji} j k t}^{-ji}, & \text{if } k \leq (K^+)^{-ji}, t \leq \ell_{x_{ji} j k}^{-ji}; \\ \frac{\ell_{x_{ji} \cdot k}^{-ji} + \eta}{\ell_{..k}^{-ji} + V\eta} r_k, & \text{if } k \leq (K^+)^{-ji}, t = \ell_{x_{ji} j k}^{-ji} + 1; \\ \frac{r_\star}{V}, & \text{if } k = (K^+)^{-ji} + 1, t = 1; \end{cases} \quad (\text{D.12})$$

and if  $k = (K^+)^{-ji} + 1$  happens, then we draw then we draw  $\beta \sim \text{Beta}(1, \gamma_0)$  and then let  $r_k = \beta r_\star$  and  $r_\star = (1 - \beta) r_\star$ .

Using the Palm formula (James, 2002; James et al., 2009; Caron et al., 2014), similar to related derivation in Zhou et al. (2016), we may further marginalize out  $G$  from (D.11),

leading to

$$\begin{aligned}
& P(\{\mathbf{b}_j, \mathbf{x}_j, \mathbf{z}_j, n_j\}_j \mid \gamma_0, c_0, \mathbf{p}) \\
&= \gamma_0^{K^+} e^{-\ln\left(\frac{c_0 - \sum_j \ln(1-p_j)}{c_0}\right)} \left\{ \prod_j p_j^{n_j} \frac{\prod_v \prod_k \left( \prod_{t=1}^{\ell_{vjk}} \Gamma(n_{vjk t}) \right)}{n_j!} \right\} \\
&\times \left\{ \prod_{k: \ell_{..k} > 0} \frac{\Gamma(\ell_{..k})}{[c_0 - \sum_j \ln(1-p_j)]^{\ell_{..k}}} \frac{\Gamma(V\eta)}{\Gamma(\ell_{..k} + V\eta)} \prod_{v=1}^V \frac{\Gamma(\ell_{v..k} + \eta)}{\Gamma(\eta)} \right\}, \tag{D.13}
\end{aligned}$$

with which we have

$$\begin{aligned}
& P(z_{ji} = k, b_{ji} = t \mid x_{ji}, \mathbf{z}^{-ji}, \mathbf{b}^{-ji}, \gamma_0, c_0) \\
&\propto \begin{cases} n_{x_{ji}jkt}^{-ji}, & \text{if } k \leq (K^+)^{-ji}, t \leq \ell_{x_{ji}jk}^{-ji}; \\ \frac{\ell_{x_{ji..k}}^{-ji} + \eta}{\ell_{..k}^{-ji} + V\eta} \frac{\ell_{..k}^{-ji}}{c_0 - \sum_j \ln(1-p_j)}, & \text{if } k \leq (K^+)^{-ji}, t = \ell_{x_{ji}jk}^{-ji} + 1; \\ \frac{1}{V} \frac{\gamma_0}{c_0 - \sum_j \ln(1-p_j)}, & \text{if } k = (K^+)^{-ji} + 1, t = 1. \end{cases} \tag{D.14}
\end{aligned}$$

We use the above equation in the collapsed Gibbs sampler for GNBP-DCMLDA.

## E Sample the Dirichlet smoothing parameter

For the hGNBP-NBFA, from (C.17), we have the likelihood for  $\{\phi_k\}$  as

$$\mathcal{L}(\{\phi_k\}) \propto \prod_k \text{Mult}(\ell_{1..k}, \dots, \ell_{V..k}; \ell_{..k}, \phi_k) \tag{E.1}$$

Marginalizing out  $\phi_k$  from (E.1), we have the likelihood for  $\eta$  as

$$\mathcal{L}(\eta) \propto \prod_k \text{DirMult}(\ell_{1..k}, \dots, \ell_{V..k}; \ell_{..k}, \eta, \dots, \eta). \tag{E.2}$$

Since the product of  $\mathcal{L}(\eta)$  and  $\prod_k \text{Beta}(q_k; \ell_{..k}, \eta V)$  can be expressed as

$$\mathcal{L}(\eta) \prod_k \text{Beta}(q_k; \ell_{..k}, \eta V) \propto \prod_k \prod_v \text{NB}(\ell_{v.k}; \eta, q_k), \quad (\text{E.3})$$

we can further apply the data augmentation technique for the NB distribution of Zhou and Carin (2015) to derive closed-form update equations for  $\eta$  as

$$\begin{aligned} (q_k | -) &\sim \text{Beta}(\ell_{..k}, V\eta), \quad (t_{vk} | -) \sim \text{CRT}(\ell_{v.k}, \eta), \\ (\eta | -) &\sim \text{Gamma}\left(a_0 + \sum_{v=1}^V \sum_{k=1}^{K^+} t_{vk}, \frac{1}{b_0 - V \sum_{k=1}^{K^+} \ln(1 - q_k)}\right) \end{aligned} \quad (\text{E.4})$$

To sample  $\eta$  for the GNB-PFA, we simply replace  $\ell_{..k}$  and  $\ell_{v.k}$  in (E.4) with  $n_{..k}$  and  $n_{v.k}$ , respectively. We note the inference of  $\eta$  for the GNB-PFA can be related to the inference of that for LDA described in Newman et al. (2009).

## F Comparisons of different sampling strategies

We first diagnose the convergence of the regular blocked Gibbs sampler in Section C.1, the collapsed Gibbs sampler in Section C.2, and the compound Poisson representation based blocked Gibbs sampler in Section 4.1.1 for the hGNBP-NBFA (Dirichlet-multinomial mixed-membership model), via the trace plots of the inferred number of active factors  $K^+$ . We set the Dirichlet smoothing parameter as  $\eta = 0.05$ , and initialize the number of factors as  $K = 0$  for the collapsed Gibbs sampler and  $K = 10$  for both the regular and compound Poisson based blocked Gibbs samplers. We also consider initializing the number of factors as  $K = 500$  for all three samplers.

As shown in Figure 6 for the PsyReview dataset, both the regular blocked Gibbs sampler and collapsed Gibbs sampler travel relatively slowly to the target distribution of the number of active factors  $K^+$ , especially when the number of factors is initialized to be large, whereas the compound Poisson based blocked Gibbs sampler travels relatively quickly to the target distribution in both cases. We have also made similar comparisons on both the JACM and

NIPS12 datasets, and the experiments on all three datasets consistently suggest that the compound Poisson representation based blocked Gibbs sampler converges the fastest in the number of inferred active factors.

We observe similar differences in convergence between the blocked Gibbs sampler, the collapsed Gibbs sampler, and the compound Poisson representation based blocked Gibbs sampler for the GNBP-DCMLDA. This is as expected since GNBP-DCMLDA can be considered as a special case of the hGNBP-NBFA, and its compound Poisson representation also allows it to eliminate the need of sampling the factor indices  $\{z_{ji}\}$ .

For the GNBP-PFA, we find that its blocked Gibbs sampler, presented in Zhou and Carin (2015) and improved in this paper to allow adaptively truncating the number of active factors in each Gibbs sampling iteration, could converge slightly faster if the number of factors is initialized to be large. However, its collapsed Gibbs sampler shown in the Appendix often converges much faster in the number of inferred active factors if the number of factors is initialized with a small value.

Therefore, to learn the factors in all the following experiments, we use the compound Poisson representation based blocked Gibbs sampler for both the hGNBP-NBFA and GNBP-NBFA, and use collapsed Gibbs sampling for the GNBP-PFA.

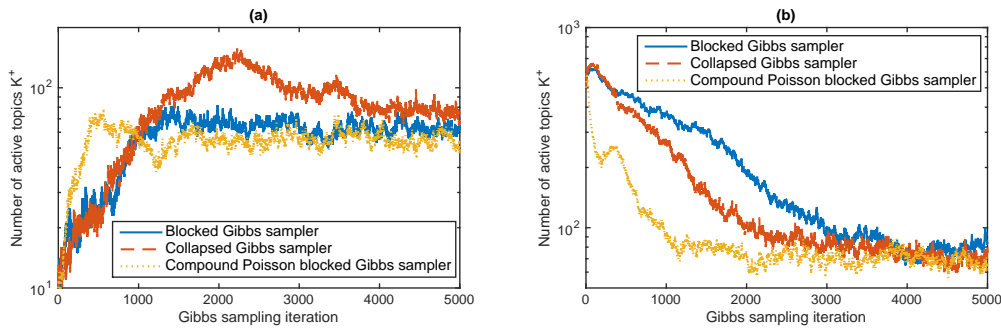


Figure 6: Comparison of three different Gibbs samplers for the hierarchical gamma-negative binomial process negative binomial factor analysis (hGNBP-NBFA) on the PsyReview dataset, with the number of factors initialized as (a)  $K = 0$  for the collapsed sampler and  $K = 10$  for both blocked samplers, and (b)  $K = 500$  for all three samplers. In each plot, the blue, red, and yellow curves correspond to the active number of factors as a function of Gibbs sampling iteration for the regular blocked Gibbs sampler, the collapsed Gibbs sampler, and the compound Poisson representation based blocked Gibbs sampler, respectively.

## G Perplexity for PsyReview and JACM

We show the results of the GNPB-NBFA, GNPB-DCMLDA, and hGNBP-NBFA on both the PsyReview and JACM datasets in the following figures.

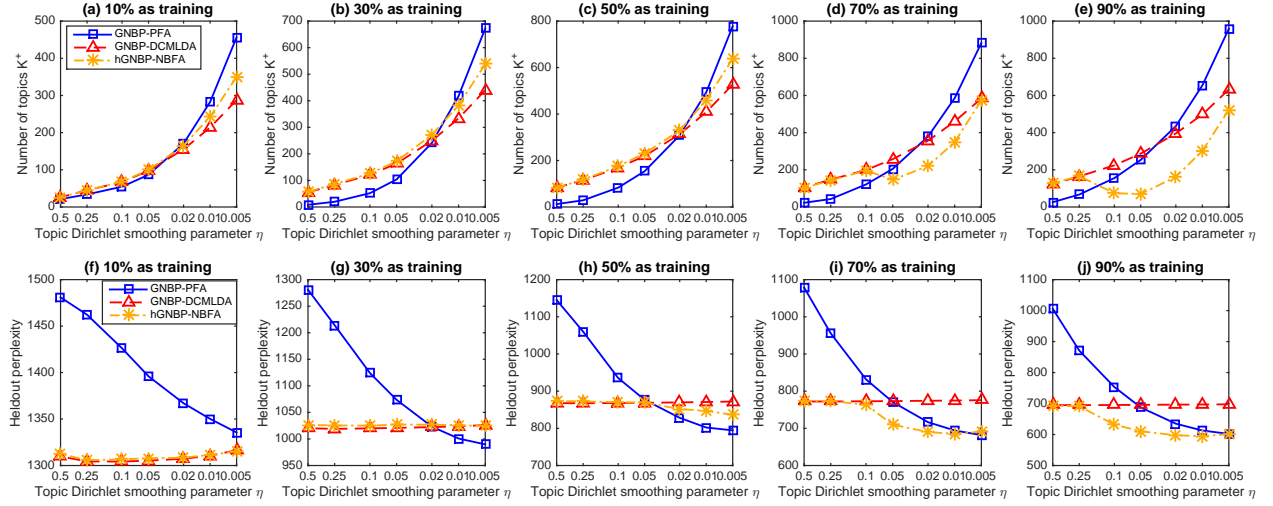


Figure 7: Analogous plots to those in Figure 1 for the PsyReview dataset.

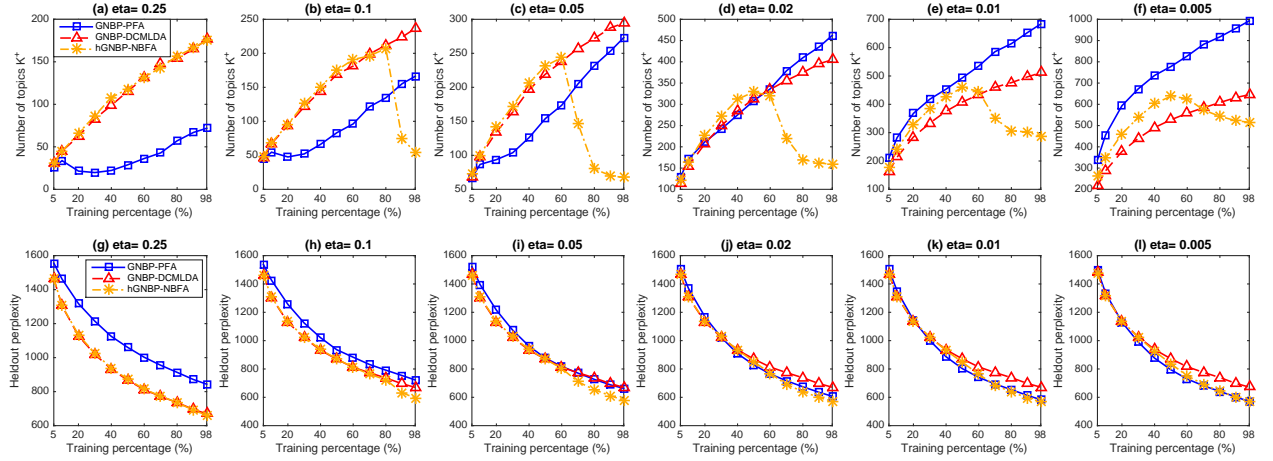


Figure 8: Analogous plots to those in Figures 3 for the PsyReview dataset.

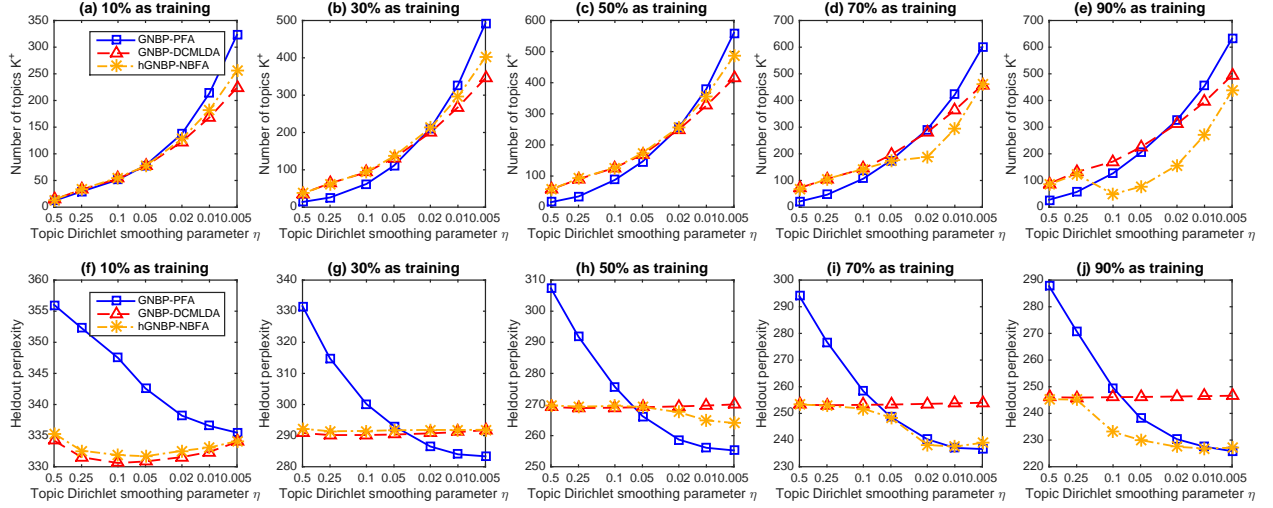


Figure 9: Analogous plots to those in Figure 1 for the JACM dataset.

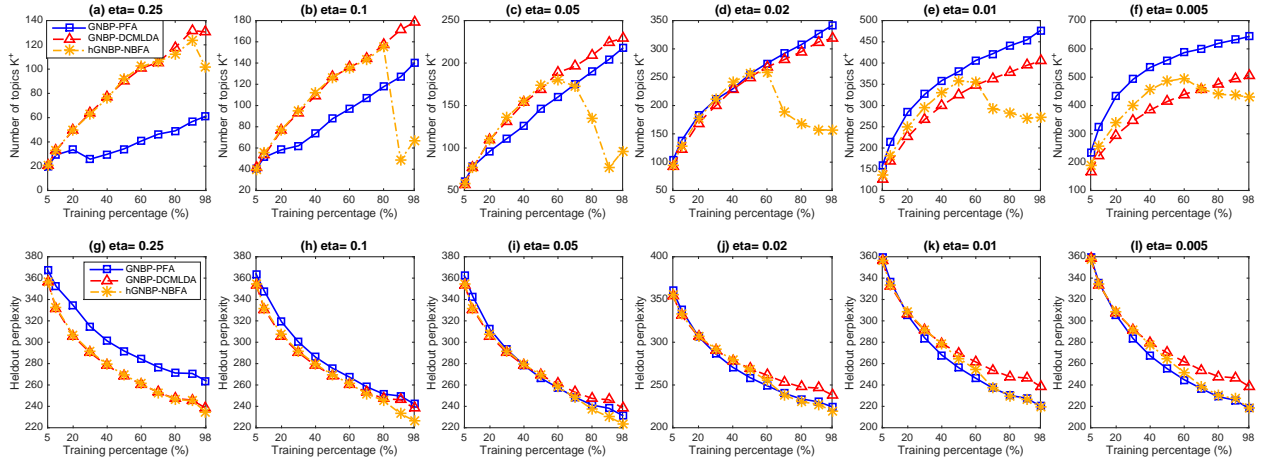


Figure 10: Analogous plots to those in Figure 3 for the JACM dataset.